GMT20200428-085107_Digital-Pr_1920x1080-converted.mp3

A [00:00:16] Ok, so, David?

D [00:00:18] OK. So just a very quick project intro. So after the initial work we'd done in-house, we got the ninety three thousand, ninety three thousand five hundred pound grant from the National Heritage Fund in January this year, and as we've kind of covered, we're working with the applied statistics and risk unit at the University of Warwick. We've got a couple of local record offices, Dorset and Gloucestershire, corporate archives from TFL and then two academic archive partners, Brotherton and the design archives, at the University of Brighton. And for those since we've got some some new people. Quick intro to the kind of risk modelling we're using. So we're looking at Bayesian networks. So the general idea is you have some sort of trigger ~~feel~~ for a risk. So taking an example after film Armageddon, know it, there's a meteor out there somewhere that's been detected that's on a collision course with Earth. So the risk event is that the meteor will actually strike Earth and that has the consequence that there'll be some loss of life. But then hopefully we can use risk management so we can try and reduce the likelihood of risk, of the risk event. So in this case, that would be trying to send Bruce Willis and his motley crew out into space to explode the meteor. Or we can also try and reduce the impact of the risk event. So in the film, they also start moving people underground so that even when they're done, if it's the meteor still hit the earth, then the actual consequence is still reduced as well. And this is the overall network we've come up with so far for digital preservation risk, the kind of two key things that the bottom two nodes. So that's when we when we've put the risks, the quantifications in that we're doing today, then that's the kind of score you'll see at the end. Based around those two nodes. So whether we can still render a file and whether we have full intellectual control over all the material as well. So we can see all of the different things that go into that, whether it's, you know what we're actually storing things on how often we're refreshing the copies. We're keeping hold of the general operating environment that we're in. So Windows and Linux and that the kind of technical policies behind that and physical disasters. Were initially really just covering flooding. Obsolescence, whether that's of the physical hardware or the the file formats and things. Technical skills that archivists and other people have got to deal with that security or all those things all linking together to give us those two final things that we can show the kind of overall rescore on. And so, yeah, just to reiterate, so, you know, you've all been involved in digital preservation, so you understand the risks, you'll have each have some different experiences. So, you know, some of these things do happen infrequently. So it may well be that one ~~one~~ organisation has experience something that another organisation hasn't. So that's what we want. It good wide range of backgrounds and, you know, give us your general genuine beliefs and experiences and give us the best overall model for everyone to then be able to use in the future.

M [00:04:12] All right, I'm unmuted that's ok. Thank you. Yes, structured expert judgement. So that's what we're here for today. You are the experts. And we want your judgements. So it's a widely used technique. The European Food Standards Agency has a very long and very tedious document put together telling you how to do it for assessing risks in foodstuffs. And it's been used for a lot of things where we don't actually have data. Things like rare events like volcanoes and things like an unchartered waters, like food prices after Brexit. Things that are too expensive to run experiments for like how do you improve pollinator abundance, given that it's very expensive to do those kinds of actual practical experiments? So, why do we need experts very often for framing and structuring the problem, for identifying the relationships between the things that are influencing it, what are the variables and how are they related? To find out where the sources of data might be

that would help and for quantifying uncertainty, because if we have a lot of data, then we find we can we can look at the variance on the data and we've got some idea of what the uncertainty will be. But obviously, if we don't have that, then we we do need to say where are the sources of uncertainty? How uncertain could this be? How low could it possibly be? How high could it possibly be? What's it most likely value? So I should just say I have got the chat up in my other screen. So if at any point I lose you or I say something which is unfamiliar and I haven't explained it properly, please do pop something in the chat and I will try and keep my eye on that. And if anything, at the end of my session, if I haven't covered that, please do, say it again. Hannah, can I don't see slides myself or will you do that? Oh, you're doing great. OK, so let's check but just basically has a three three phases. So the pre elicitation stage is where we define the problem. So the problem in our case is digital preservation. We don't know what strategies are the best strategies to maximise the probability of our digital preservation, given the constraints that we have. So we've been hunting for experts and we have found you and we're very delighted to find transition a group of high quality experts. And I'm doing the facilitation today. I've done this before. The slides by [00:06:48]? [0.0s] and Hannah, who devised the protocol that we're going to use. And she basically trained me how to do this. Validation data. So is what data is out there? And then frame the problem. So what is it we're trying to do in this case, what we can do is make a decision support system. The middle box is the elicitation box, that's what we're doing today. And then afterwards, we're going to be doing some statistical analysis on the judgements that you give us. And we are going to then incorporate that into the model that we build. Next slide.

[00:07:22] So how do you do it? Well, nobody knows the best way to do it, because in different circumstances, different protocols perform better. But we do know that the more experts we have, the better. And 20 is a brilliant number to have a lot more than that. A lot more than that. If you have 30 or more, you probably aren't going to add any value if you have five or fewer. You're unlikely to get the kind of range of ideas that you want for a really robust elicitation. So the fact that you've all suggested a [00:07:50]? [0.0s] is making us extremely happy because we're just on the sweet spot there. Preparation and planning, which put a lot of time into that, is important in how we phrase the questions, how we sequence the questions will affect to some degree the answers we get. So we've been very careful to look at that hard and try and make sure that it's as unbiased as possible. And then we're going to aggregate multiple judgements. So the numbers you give us today are going to be pulled together in a mathematical sense and tells us. And I will be doing that overnight. So we'll give you some results to tomorrow. And we're documenting the process very carefully so that if in some point in the future, we need to go back and say, well, why should we do it that way? Or why was this defined in this way? We have got a record of that.

[00:08:37] OK. So, thank you.

[00:08:40] So we do need a diversity of opinion, so we could do we could have done this exercise simply with the stuff at the TNA and that would have been great. They'll definitely fall into that into the category of experts, but they are working at TNA and maybe they don't have that much experience of others or maybe they do. Depends on the career path, of course. But we need this diversity of opinion and it's really important that we actually collate, just like you would in in in some experimental data. You'd probably have a bit of an outlier here and a bit of an outlier there. And those are the ones you look carefully at. You can throw them away. You would say, oh, maybe there's some aspect of the story that this outlier is telling us. So, again, we [00:09:22]? [0.0s] and Independents who well, we don't want anybody to say, well, actually, we think William knows everything about this. So we

will just tell you the same thing as William would have told you. We really want your you to think independently. How is your career developed? What knowledge of you gained along the way? How, what things have you seen? and give us your honest opinion about that. And again, the decentralisation, this is this is that everything can't be coming from London. It's gotta be coming from other places as well. And the aggregation, again, we're going to average in some way the responses we get and get a distribution, a range of values for that which captures all the expertise that you bring to us.

[00:10:13] Next slide, Thank you.

[00:10:15] So, yes, a lot of experimentation has been done about experts and how you ask questions and how you of fix the answers. Those of us who are old enough to remember. Yes, prime minister. And how Sir Humphrey says well the prime minster says we'd like like like a survey. It's obvious as well. What answer would you like? And he goes on to demonstrate that how you answer. Ask a question does influence, in fact, the answer you get. Making judgements under uncertainty is not easy. You can expect to be tired after today. Thinking about all the different things which might influence the value you're going to present isn't an easy, easy process. And most people are not great at judging probabilities. So another way you can do that is to think about a thousand of something or one hundred thousands of something and turn it into natural frequencies, which some people find easier. And then we all have we will use [00:11:09]?. [0.0s] We will have biases, you know. You might think very you know, our background by by nature gives us some biases towards something in a way from something else. We'd need a lot more evidence to move our opinion from where it is to believe that things and to believe that thing. So we try in the process to reduce those as far as we can and to identify those.

[00:11:33] The next slide, please.

[00:11:36] So where are we going to guard against? Anchoring bias.

[00:11:41] So the first piece of information that you say will, influence where the negotiation goes. So this is a salary negotiation. Who makes the first offer that puts the ballpark of what's possible? Some of these offer you 20000. You don't then go and say, well, I want a hundred thousand twenty two, but it is not usual. OK. So some people are angered on some piece of information that they heard on the radio on the way in or something like that. So the way we ask questions tries to reduce that bias. Next, Availability. So people think, well, I've seen this a lot. So this is really, really important. So in this particular example, smoking doesn't cause cancer at a huge rate because at least two of my granddad smoked 100 a day. And they're fine. But actually, that's not actually accurate if you're looking over the whole of the population. So, yeah, our locality can give us some availability heuristics. What else have. Overconfidence. Yes. So I've got this. This [00:12:52]? [0.0s] doctor in front of my name. So I must be an expert in everything, right? So academics are prone to be overconfident because what we just saw. And so this has to do with credentialing as well. So people you think of as experts of people with the high profile from the big names and all this kind of thing. But actually, there's a really nice story that Mark Boeckmann tells of an illustration he did. On geopolitical events, and they had an admiral who was part of the expert panel. But they also had somebody who was a secretary in a mining company. And it was really interesting because although the admiral had been doing geopolitical forecasting all of his professional life. The the the woman who he probably would have called trailer trash left school early. Lots of children, secretary in a mining company. As I said, she was just really, really interested in geopolitical events. And so she would listen to the radio and she'd come and talk to it and everything like that.

When she was taking part in the in the elicitation when there was a question come up about a particular country should go and talk to somebody in the in the organisation that she knew who came to that country to explain about their political system. And she would make her judgements. And and so she actually scored at least as well as the admiral. And in some cases, better. So it's not always easy to see from credentials who might be a good expert. But we believe that you're all very good experts. So we thank you for coming today.

[00:14:31] Yes.

[00:14:32] So we need to guard against groupthink. You know, it's very easy. It would be very easy for the young lady concerned to agree with the admiral all the time, because obviously he knows what he's talking about. That doesn't seem right to me. That's not what my opinion wasn't my opinion. But I will change my opinion. So we need we need people to say what they really think. We don't want one member coming out of 20 people. We on to a range of ideas and the next.

[00:15:03] And we don't want to discount somebodies opinion because of some characteristic that they have.

[00:15:10] Next.

[00:15:13] So aggregation is the idea that we're going to take all of your values that you provide for us and come up with a single number or range of numbers to represent the knowledge that all of your group has. So, again, the divergence of opinions is on how to do this. Sometimes people. Go round and discuss and come to a sort of consensus opinion, which inevitably means some people are having to move away from the opinions [00:15:40]?. [0.0s] And another way is mathematically. So sort of taking an average or weighted average. That kind of an idea. And so we're going to do in this political a slight mixture of those. We are going to have a discussion that there might be a chance of two people changing their mind because if they extra information. But then we get do a mathematical aggregation of the of the numbers.

[00:16:08] So we need you to understand the questions we need to have questions with clear operational meanings. We need a methodology that's clear and well set out. This is not something we've invented ourselves. It's well established. It's been tested and proven. We need to try and make sure that we remove as much of the bias from the data as we possibly can, just as we would in designing an experiment. So we know who you are and we will know on the out on the private copy who said what. But in public, obviously, its Chatham House rules. What was said can be reported but not who said it? So the empirical control is about the removing of the biases wherever possible. And we will keep a proper record of everything that we've done today so that we can go back to it. Should that be necessary.

[00:17:02] So, so here you have to write in the chat, everybody. So what percentage of a watermelon is water? So you'll notice that we've got the lowest plausible bounds, the upper plausible bound and the best estimate. The reason we have it this way round is to reduce anchoring. So if we have a if we give our best guess to 50 percent. First, we tend to prove plus or minus the same amount. But we think about how low could it be? How high could it be? And then what is the best estimate? Then we reduce that bias. So. We know that a watermelon is called a watermelon. So we might think that the water content is quite high. We know it's not 100 percent because otherwise it would be a droplet of water. It does have skin. And it's certainly not zero because it would be dust. So it's not a physical

round. So the five percent here is how low could it plausibly be such that you would be really surprised if it was lower? Well, it's absolutely impossible to be lower, but you'd be really, really surprised if it was lower. So I might think that a watermelon is 50 percent water and I'd be surprised if it's lower than seven percent water. It could possibly be, but I think it most unlikely or I might think it's it's 50 percent water. It could be as much as seventy two percent water. I'd be very surprised if it was high, isn't that I mean, how could it hold together? That might be male reasoning. And how I think about that, though, I have a quick go at that. So three numbers. First of all, five percent, second to 95 percent. And thirdly, 50 percent.

[00:20:09] I'm typing these quickly into a spreadsheet.

[00:21:08] Williams confused me by putting in the [00:21:10]? [0.0s] a page or two.

[00:22:55] Excellent.

[00:23:21] Okay.

[00:23:25] Okay, Jodi, good got percent. So what do we what do we come up with between ourselves? So the average of the lowest plausible band was 25 percent water and the average of the upper band was eighty one percent water. And the average of the best estimate was fifty nine percent water. OK. So actually the answer is 92 percent by volume. So I think what I did that by suggesting to you some a value of around 70. I think I gave you a recency bias. I made you think that it was actually going to be lower than it really was. OK. So you can see, though, that between us, we got nearer to that, to the truth than a lot of us did individually. So we've actually done that rather well here. So by buying by rather and asking just one person. And we've actually gotten over to the truth by asking lots of people.

[00:24:26] So thank you for that.

[00:24:44] OK. So how do we do the empirical control? Loving these loving these answers. So how do we do the empirical control? What we do is we have a series of questions for which the answer will become known soon. But it's not known to the experts currently or perhaps is part of some information that we have which isn't public. And we ask them questions of a similar kind to the questions that we don't know the answer to the questions of interest. And what we do then is we were able to see how accurate and how informative individual experts are on the people. That. Are more accurate and more informative. They're slightly up weighted compared to the people who are less accurate and less informative, but everybody contributes to the answer. OK. So what we can do then is use the user scores on those as weights.

[00:25:52] Next slide

[00:25:56] And as I mentioned before, with the with the lady from the mining company, it's actually very difficult to tell the difference between an expert and not an expert. Credentialling doesn't always do it, but we're fairly confident that we in the in today's game, if all we can say, do you want to be experts? So we're looking forward to a good result here. I'm not an expert on watermelon, obviously.

[00:26:25] So these calibration questions again. So this is this. We will ask you to assess your uncertainty about some variables and some of them will be calibration questions

when we have access to or we'll have access to the true values. And we got to be able to use those then as a weighting scheme. The assumption we are making here is the future performance of experts on variables of interest can be judged on the basis of their past performance on the on the variables which are called sea variables or calibration questions. So that's the idea. So we're asking you questions that are within your range of expertise and you do well on those. And we expect that your answers are for things that none of us know will also be really good.

[00:27:07] Ok, next slide.And this is what we would get then. So you can see there are calibration questions going from the bottom on the left hand access question one, two, three, four, five, six, seven, eight, nine and 10 along the bottom, there are some values.

[00:27:24] And then the colour coding is four different experts and are, which is the right answer. So these are calibration questions. And we can see going to the top of the graph. I'm looking at question time and I'm looking at the Red Range graph. We're looking at. So that is expert D. I can see that was expert D, didn't get the right at his or her best estimate was not exactly on the right answer. They did. They did capture it within their range. So that's a very good accuracy. And we also can see that the length of the distance between the five percent, the lowest plausible and the 95 percent the highest plausible is relatively small. OK. So this this person is both accurate and informative. Now, experts see the green one also captures the right answer. But their range is much, much higher, which either means that that expert thinks that the underlying system is very uncertain or they have personal uncertainty about where the value should be. So that person is accurate, but not very informative. So then the Blue one, which is expert B can see that they've just about missed the real value is lower than that. That's what they consider to be the lowest plausible value. Reasonably narrow. Uncertainty bands. So, so not accurate, but quite informative. And then you see the expert A, very, very narrow and certainly been very, very lucky to be overconfident because we can see this pattern going all the way down. The pink is very, very narrow. This is the real value. So that person is is informative but very inaccurate. So misses by quite a lot. A while. Quite a bit. And not consistently either. So question nine. They were to low. Question eight, they were too low. Question seven they were too low. And they are consistently, consistently too low all the way along by either optimistic or pessimistic, not managed to capture the real answer any time. So for our aggregation on this kind of performance to be uprating experts C and D with D having slightly higher weight, but we did B down weighting the answers given by B and A.

[00:30:00] I hope that's clear.

[00:30:03] Yes, good. So the idea protocol, which is what we're going to use today, which is develop by anchor Hannah and her colleagues. We've done that defining the problem, the pre elicitation, and now we're in the elicitation. So how does that work? So, first of all, first set of individual estimates. So investigation, you've done that in some deep into the previous workshops. You had the information ahead of time about what the model is. So you've had a chance to think about some of things that you might be asked about today. And what we're going to ask you to do is without consulting other experts, just to give us your best estimates of the questions we can ask you. And they're going to be exactly in the same format as the as the watermelon one. We're going to ask you for the lowest plausible, then the highest plausible. And then your best estimate. Then what we're going to do is give you some feedback and have a facilitated discussion. So the range graphs we saw for the calibration questions on the previous slide. We're going to have some graphs for going to work overnight. Get those graphs ready from the information that you give us. We're going to have a slide for every every individual question. Are we going to have a

discussion about under what circumstances? This particular. The answer to this question could be as low as this or as high as that. Just so that we feel like we're all on the same page in terms of , in terms of information. So if somebody is read a research paper that came out last week and is able to feed into that particular discussion on that particular question, and you haven't had a chance to read it yet, then that gives you a chance to think of actually that does actually feed some useful information. And then we'll ask you tomorrow to give us a second set of individual estimates. And again, these are going to be individual private estimates. Nobody will know except us what you've written down. So if you if you're waiting for promotion, buy from somebody on the call, you don't actually have to then publicly disagree with them and put that in jeopardy. There's no kind of social pressure to give an opinion, anything other than your own. And then what [00:32:06]? [0.0s] I will do will go away and we'll aggreagate the expert judgements and we'll inform and insert them into the model. And this will become part of the decision making. Hopefully archives everywhere.

[00:32:18] OK, next slide.

[00:32:24] So, again, to repeat, the individual assessment avoids anchoring on other people's estimates we're seeing from the, I think the melon that I have managed to anchor you on this near 70 than the 93. Discussion helps with the availability. The best thing that you think of, best thing that comes to mind. I know liberty is the second ground reduces a group think and the dominating effects and the way we've designed the questions helps to reduce the anchoring and the overconfidence.

[00:32:55] Next slide.

[00:32:57] Yes.

[00:32:57] So feedback and facilitated facilitated interaction have been shown to actually improve the end result. So when we do the calibration questions, we know that the second round after discussion, almost everybody goes closer to the actual real answer. So we are assuming, of course, that by doing that with the ones that we don't have the answers to. There is a weak dependence that's induced between that. What we try to minimise that by allowing you to do the second round in in a private way as well. And then we we. We will. We will do the mathematical obligation, depending on the weighting and sometimes equal weighting is the best way to go. And we can test that. So that's when we we will we will do that kind of behind the scenes and work out how to best segregate things.

[00:33:54] Next slide.

[00:33:54] OK. So this is a picture of where we're at. So this is the previous before the draught, the calibration of listeners who questions who done that? Identify the experts and provide background on that. OK, so the next phase for today is we are going to go through the questions and we go to check that they are clear. So we can write down a question in a particular way. It's clear to us what we mean, but we might give it to three different experts and they might in their heads be answering three different questions because they haven't quite grasped we haven't written in the way that's foolproof, shall we say. And so we've got to go through the questions and just check that everybody's understanding. The same question from the ones that we've done our best. But we are only human. So then we're going to ask you to make the initial estimates and then we're going to discuss this. Some businesses similarities and differences of opinion. That's

tomorrow and providing a final confidential estimate. And then again afterwards, begin to give, give, give feedback and do the analysis and the aggregation.

[00:34:59] So we do need you. As I've said multiple times, we need the diversity of your opinions. We do need people who have had different experiences. People who are handling different kinds of data. People who are having different roles within the preservation cycle because it gives us different viewpoints. It gives us a better viewpoint. We can't work out how wheat is going to grow by planting in one field, one on soil type. But we can't do a good aggregation with people who all think the same. So please give us your genuine beliefs, even if you disagree with other people. If you really think that after you've heard that their reasoning for that tomorrow morning, you really think that you beg to differ. Then please write your genuine beliefs down, because that way you really help us.

[00:35:49] I haven't seen any questions coming upon the chat, are there any questions that people want to ask me now, have I? Is there anything I haven't made clear?

H [00:36:04] Please feel free to either ask questions on the chat or or, if you like, on mute yourself and ask questions now we want to make sure you get the opportunity to be as fully informed as possible. So please don't be shy if you do want anything clarified.

M [00:36:21] Thanks (name), thanks (name).

M [00:36:30] Yes, we will need to share definitions of terms that we are using so on the question sheet, Hannah has worked very hard to put definition boxes. So we will go through those as well. Make sure that those definitions are workable. Thank you, (name), for that question.

H [00:36:50] And along those lines, you won't have anything from me yet. But during this morning break, I am going to e-mail you all with the list of questions and the answer sheet and the helpful definition page. So that will be hopefully in your inbox by the time we rejoin at 11:00.

E20 [00:37:06] Can I ask a question, Martine, which is a bit about a little bit around the configuration of digital preservation or the kind of contitution so hear me out. There are different organisations on the call. That's a good thing. So business archives as well as universities, as well as the National Archive. And that's a good range of digital preservation. But there will be, I guess, subtle differences of answers relating not just expertise, but the context in which that answered. So the experience. For example, (name) in a local authority, may be different from the experience of, I don't know, Alex at the National Archives, because close context is different, they would have a very honest, straightforward, potentially quite precise and accurate answer. But the context or deference, make, throw something up. I wonder to what extent, therefore, you need us to answer a generic way or do you need just to us answer based on our very kinda closely defined experience.

M [00:38:06] Well, so the questions are very much raised. What would you expect? So we have a list of the experts who will know where they've come from and what we're trying to do in the phase one of the model is to make the model as widely applicable as we can to want to catch or capture the range. What we would like to do if we get the opportunity to develop a phase two of the model is to make it slightly customisable. So the different bits are in there because, because of the TNA that are not relevant to your particular archive. You can then knock them out. That's what we would like to do. But for today we would like

you to answer for yourself because that will then provide us with the data to do the customisation as well.

E20 [00:38:56] That's very clear. Thank you Martine.

M [00:39:09] Any more questions?

M [00:39:20] Well if anything, comes into your head? Do feel free to say so, we'll pop it in the chat room until we can. We can answer it then. We're a little ahead of schedule. That's good, isn't it?

H [00:39:33] Yes. So in that case, if if no one's got anything else to ask now, if anything does come into your head. Do pop in the chat or come back to the call. But in that case, I guess it's time for a break and we will try and stick to schedule in case some of you had plans or calls. So if you could return the call to start again at 11:00 p.m., we'll start discussing the question. So between now, it's about half ten now and we'll come back on at 11:00. Probably make one minute past 11:00 because as Alex said, there is a minute of silence and so half an hour go make yourself a cup of tea, do your other calls and any other work. And we'll see you back on this same call at 11:00. Feel free to just stay on mute and go off and come back again, whichever you please.

M [00:40:19] Thank you Hannah.

H [00:40:20] Great, see you at 11:00.

H [00:40:58] ~~What the plan is.~~

H [00:41:00] For the next, thank you for pressing record, I assume that's Alex. For the next for the next two hours and hopefully, hopefully went well beyond that. We're going to focus on looking through the elicitation questions and checking we all have a common understanding of them. So I am going to change my screen. So bear with me and I will get them up.

[00:41:23] ~~And. You share.~~

[00:41:37] OK, so.

H [00:41:39] Hopefully, you can all see a new slide that says Digital preservation risk model elicitation questions. Oh, it says one and one, right, that should say one and two. And first and pause there. And Martin, are you on the line?

M [00:42:01] I am.

H [00:42:02] Excellent. So do we go through this one by one. Have you got any advice?

M [00:42:10] Yeah. So I think. So I think we can. The best thing to do is to just say this is this is question one. Is that clear to you? And then any got any questions can come on and say what do you mean this or do you mean it like that. And then you can answer the questions. And when somebody is happy we move on to the next question.

H [00:42:28] OK. So question one, which hopefully you can all see on the slides, but you should also have on the question list that has been sent round and on the answer sheet

that's been sent round. So hopefully you'll have three copies of seeing this. Sam you're on the phone and you're not missing anything. It's we're looking at question one on the question list. We're looking at on the slide.

Question 1

H [00:42:54] So out of a thousand U.K. archivists, how many would you expect to say that their digital archive has a list of all the digital file formats collected? Does that make sense to everybody? Does anyone not understand what that might mean?

E07 [00:43:11] Is that, does that mean they've got a list of all the file formats in the collection?

H [00:43:19] Yes.

E07 [00:43:20] Yeah.

E20 [00:43:24] And does it mean that all of the digital file formats or are you expecting that some of them will have some uncertainty? Even though they may have a list.

H [00:43:43] So I would interpret this as though they have a list of everything that they have in their collection. They have a list of what those file format are, where possible, I guess if here are file formats they don't know..

D [00:43:58] Yeah there will be formats that for various reasons won't be you know, can't be identified absolutely. That's fine, I think.

[00:44:05] Yeah.

E09 [00:44:08] Would we be expecting for this that they would know their sort of pronom I.D., or would it be sort of that they just have a more generic understanding or which they have?

A [00:44:24] I think it's more generic. We don't expect to know the greatest detail. It's just to have an idea.

E13 [00:44:34] Hi, it's (name) here. Can I just ask for some clarification on that first question? Might seem a bit odd to be my way, my mind works. [aside: just got to clear it please]. You're asking about a list of all the digital file formats collected. I don't have a good understanding of what a list means. Is that a like a consolidated itemisation or is it the distributed set of information across a number of files, for example, that can be collated?

A [00:45:24] Think it could be either, really, whatever you consider. It's about being in control, it's about knowing what you've got. So, I think the list could be something you either have got in one place or, you know you could get.

E13 [00:45:39] Thank you, that answers that perfectly. Thank you very much.

E07 [00:45:50] Can I just check? Are we meant to be answering this question. Or are we just discussing it?

H [00:45:57] No, we're we're just discussing for now. So don't worry about answering it. You don't even need to think about what. You can think about what you want might be if you like, we're just making sure we all understand it.

E07 [00:46:07] Okay, thanks.

E01 [00:46:09] Can I, (name) here from (name). the thousand UK archivists, are we assuming that there are a thousand UK archivists? That would really you help me. Is that sort of, are there a thousand repositries or are there a thousand of us. It would just really help me to understand what we're thinking of here.

H [00:46:31] Yep. So we're, all of these questions are going to be out of the thousand. So that's sort of a consistency for the answers. And I'm not sure if there are 1000 UK archivists. Hopefully there are.

[00:46:42] I looked this up actually not long ago for something else. And I think there were about two and a half thousand. And that was about 10 years ago. That was the only figure I could find. It was about 10 years old. So.

E20 [00:46:56] And asked about how broadly or how precisely thats defined, you know, because there'll be people in working in museums, you know, next sector along, so to speak, that will have a similar sort of challenge. Do we want to think about including them or is it again really talking about archivist's in a sort of archival professionals sense?

M [00:47:20] The reason that the question is phrased the way it is, is because probabilities are very hard to think about. OK, so what's the probability that an archive has a list of all legitimate file formats collected? It's really a really difficult way to think about it, whereas if we say out of a thousand U.K. archivists, how many would you expect to say their archive has a list of all the digital files formats? We're essentially asking the same question. We're actually asking you for a number of natural frequency out of a thousand. We've done that for consistency all the way through. So, it really isn't, It really isn't about the thousand, it's really about, this is a way of framing the question which makes it, brings up a mental picture.

E04 [00:48:12] Hi there, can i just clarify my answers, do you insert a number slice of a thousand rather than a percentage?

M [00:48:22] Yes.

E04 [00:48:23] OK, so we're dealing with thousands number, ok.

H [00:48:34] Any more, clarifications about question one? As we go through quite a lot of them start with out of a thousand UK archivists as well.

H [00:48:46] OK.

Question 2

H [00:48:56] So, in that case, I can't hear any more questions, but do shout or comment. If so, question two, though, on the slides it looks like question one again. Again, out of a thousand UK archivist's, how many would you expect to say they that they had at least some knowledge or skill to perform file format analysis of a digital accession? Is that clear?

E07 [00:49:28] Is that the same as file format identification?

E07 [00:49:38] That's my interpretation of the question, but analysis I would see, possibly being something like sort of further on, like the next step from file format identification. It's more about looking at the results of that identification and, you know, running some reports and and graphs and thinking about, thinking about what they've got more.

H [00:50:06] David or Alex, do you have any thoughts about how this analysis should be interpreted?

A [00:50:14] I though we were thinking about it in quite a simple way in that it was just looking at finding a tool that would be able to tell you what file formats you had. I would take it as that to begin with. I think.

E07 [00:50:29] OK, thanks.

A [00:50:32] Yeah, it's definitely not appraisal its around, yeah ok, so identification.

H [00:50:40] OK, so think about that as identification.

E10 [00:50:50] Yes, it's (name) here from (name), is that using a tool that's not just looking at it in your Windows Explorer and identifying a JPEG file that's different from a word document? Does actually, does involve does it involve the formal tool type today? A: Yes.

E10 [00:51:17] Thank you.

E05 [00:51:26] Hi, sorry. This is about the scoring rather than the specific question. But just because I'm confused, are we doing it, When we get this going, Is it like the watermelon one where it doesn't have to add up? It wasn't adding up to 100 percent. So, like, each percentile is out of a thousand. Or is it that overall your answers have to add up to a thousand for each of the answers?

M [00:51:46] Each one is out of a thousand. So you can't have more than one.

A [00:51:55] You're a bit faint, Martine.

M [00:52:00] Is that better?

[00:52:01] Yes.

M [00:52:04] So, It's a number out of a thousand. You'd be surprised if it is lower than, say, 10 out of a thousand, and you'd be really surprised if it was as high as 950 out of a thousand. But you probably would think it's something like. Some 700, 720 out of a thousand would be about what you would expect. So that's the kind of answer we're looking for. Not to not wishing to anchor you on a particular numbers.

M [00:52:31] So, again, the lowest plausible. You'd be very surprised if it is lower, The highest plausible, you'd be very surprised if it was higher. The most likely answer, in your opinion.

A [00:52:42] Definitely doesn't have to add up to a thousand then

H [00:52:48] Alex, have you seen (name) message in the chat? I.

A [00:52:55] [read's question to self] I think either, it could be either.

H [00:53:02] Yep. So do I.

A [00:53:05] Yeah.

H [00:53:10] So they're not totally bewildered, they are going to know somewhere to start even if they've not done it.

A [00:53:16] Yeah.

H [00:53:24] Any more questions about question two?

[00:53:32] Hi. Sorry. Question two. Because we're right at the beginning of this, I'm assuming that this is kind of going to get a bit more granular later on because there is a big difference between having the knowledge and having the skill in terms of the risk effect.

[00:53:56] Yes, so. So what we're doing here is we're plugging the gaps in the data that we have. So we've asked these questions very specifically because we have some data, but we don't have enough. And so the way we ask these questions is particularly to to plug the data gaps that we have. So we will be combining the answers to these questions with with data we have from other sources in order to make the risk. Sure overall.

E13 [00:54:22] Can I make an observation here? Martine, its (name) here.

E13 [00:54:28] I I've got the questions printed out for me here And I glanced ahead. What I'm finding is that I'm having to adjust my thinking on on, for example, question one, by starting to think about question two. Because otherwise, my numbers don't seem to make any sense. So, what my observation is, is that the, is that the order in which these questions are posed, actually has an impact upon how one thinks about them and the belief that one might have.

M [00:55:15] Yes, and we've been very conscious of that, would you be putting it together. By going all of the questions before anybody starts to try and quantify, then you will have a good overview of what it is we're asking and we'll have had these conversations. And so everyone's head will be in the right space. We're hoping. That's how it usually works.

E13 [00:55:34] Right. So. So at the moment, then, you don't want us to actually provide the answers. You just want us to understand what the questions are?

M [00:55:41] Exactly. That's what you're doing now. If you want to write down something because it's coming into your mind right now, then feel free to do that. You can re-visited it when we've been through, all the questions and check that you still think that's the answer you would like to give.

E13 [00:55:53] Thank you.

M [00:55:54] Welcome.

H [00:56:06] OK, are we ready for question three?

Question 3

H [00:56:12] OK. So there is some definitions here as well on the screen and on your piece of paper, you might want to familiarise yourself with if you're not aware of some of these terms. But question three is: out of a thousand U.K. archivists, how many would you expect to say that there was some capability within their organisation to carry out at least one of A. File format migration, b. Software emulation or C. Data recovery from damaged or obsolete media, even in a limited way? So some capability within their organisation, not necessarily the archivist, themselves, and to carry out at least one of those three skills.

E10 [00:57:13] Hi, it's (name) again from (please). In defining capability. Is that the knowledge? Of how you would go about doing it or it could might be done? Or is that the technical capacity of the organisation to be able to do it?

A [00:57:36] I think we were, we did talk about this, didn't we, I'm trying to remember now. I think we were talking about, you know, can you actually do it? And somebody in the organisation, Do it, you know, so. Because if you think about the risk being whether you kind of when you can't, it's about whether somebody can actually do one of those things.

E13 [00:57:57] It might be that you know the telephone number of the [too faint to hear] to do it.

A [00:58:07] Does that help at all?

E10 [00:58:08] Yes. Thanks.

H [00:58:14] I just would say [00:58:15]we've also got a question: [0.0s] Does that include having the funding to send damaged things to external experts? So I guess this is capability within your organisation to form one of these tools. So the question is, are we including outsourcing tools?

A [00:58:37] Well, I suppose that mitigates risk, doesn't it? If you can, if you have the capacity and resources to send it out, to send it out, then that's you being able to mitigate the risk of those things being a problem. So I think yes, I think? it does include that?

D [00:58:55] And you at least know enough, enough then to actually seek out a suitable.

A [00:59:00] Yes. I know what you're asking. Yes.

A [00:59:11] Yes, does that help? yes, similar to hiring a contractor. Yeah. Getting somebody else, somebody in to to do something specific. [reads comment]

A [00:59:26] No, that's true, but the idea of this, I suppose, and anybody can step in and correct me if I'm wrong, the idea is that can you organisation do something about this to mitigate the risks? That's really where we're going with this. That's what we want to be able to say you cannot you can't do.

A [00:59:46] If I was working a small organisation that didn't have a great deal of funding, then I would either have to work it out myself or I would just have to find another way of mitigating that risk. Does that make sense?

H [01:00:10] OK. OK. Any more questions that people are familiar with the terms, file format migration and emulation?

H [01:00:29] Right. great. Right. We'll move on to question four. So we have got some terms here that might want to be careful with. So the question is: Out of 1000 files, with file formats, that are ubiquitous and/or open, How many would you expect to have the tools to render? And a few definitions that are useful here. So ubiquitous, we're saying widely known and used by non-specialist. So I would give an example that Microsoft Word is ubiquitous. Lots of people use word. We know how to use it. And then we have open. So by open, we mean, not proprietary. So something that ultimately, even if it wasn't widely used anymore, you could perhaps write software yourself or someone would have, would have would have built some tools to render it because the file formats open and expect to have the tools to render. So this is having the available tools. And software to render this. So even if it's word, do you have word on your computer, can you open it? So most peoples would be. Yes.

E20 [01:01:57] So can I ask a question on this one? Because my sense is it would be often times possible to open a file to render a file. But the question arises about being able to open it to a degree of satisfaction with a degree of confidence about the performance of the file, when open. So, an audio file, for example, with sound dropout, a video over the sound thats out of sync. Theoretically, they're open, but of course they're not open to a satisfactory level. Equally, word process documents with the font's have been replaced. It's possible to open the file, but the font has been replaced and consequently some element of the performance of the files been lost. Could you get your to think a little or explain what you mean in that context?

A [01:02:49] Think with tools to render. We were hoping, I think because we don't want to. I didn't want to go down the significant properties, rabbit hole. So the idea really is, is that whoever is looking at it decides that it's good enough for their purposes. So if the fontsare different. But the layout of the page and information is all intact, Maybe that's that's fine. Whereas a video where the audio is dropping out is obviously not really of any use, you're, you're losing something that's that is a fundamental part of that file. So I think that's what we were talking about. I think thats what we we talking about when we were thinking about rendering in that sense, it has to be good enough.

E20 [01:03:36] Yeah.

Question 4

H [01:03:38] So this this question is particularly about do you have the tools to be able to render something? We do have a question later about even if you have the tools, do you know your rendering it properly? Because I know that something with digital material. OK. We have a word doc, and we have word. But are we, do we know we're doing a faithful kind of real render of it should be, so that it's something else which we'll look at later. This is focussing on do you have the tools?

E13 [01:04:11] Hi, it's (name) again, can I ask another question? This is based on experience I've suffered. Your question says. How many would you expect to have the

tools to render? This is subtly different from how many would you expect to know that they've got the tools to render? I've experienced cases where I've been asked to render a an a supposedly strange file format. And have actually shown the user that they already have the tools on their machine to do it. They just didn't know.

H [01:05:01] That is a good point. So, thank you.

A [01:05:07] I suppose that's around expertise, I mean we would know, we would expect it will because were saying out of a thousand files and file formats, and suppose again, we are talking about UK archivists who will therefore know a bit more about how to access files and what software they could use to answer things.

D [01:05:23] Yeah, our definition of tools to render says and the expertise to use them so.

E13 [01:05:30] Oh, right. Yeah. Yeah. I'm just referring to question four. I haven't looked at the definition. OK, thank you.

H [01:05:42] And just as as a reminder, when you do come to answer these questions, you are going to be giving the five percentile, the 90th percentile and the 50th. So we will be incorporating a range because you might be very uncertain. Exactly. You know, what the actual answer is. And that is important. We can capture that range.

E14 [01:06:05] And looking at looking at the things that are in chat at the moment. (name) has asked about, are we talking about now or in the future? I think for them for this question is about what you're holding now. Could you actually render them?

H [01:06:16] Yes.

D [01:06:16] Now. And yeah. (name). I think that goes to your the question you're asking and it goes to that sort of level of expertise, how, if you if you know where you could download stuff and, you know, it would have to be confident going off and doing that and have access to different types of, environment then it's obviously, you know, you've got a higher level of expertise. So.

D [01:06:50] That's partly, I guess, where the range comes in as well. Some people will be able to do that. More, more confidently than others.

H [01:07:02] In terms of whether this is within your organisation or outside, I would like to say within, because this is the risk for your archive But I don't know gain if this is the kind of outsourcing question. Alex or David, do you have a view on wether we should be answering this just within? [read comment] Yeah. I think I agree (name).

A [01:07:24] Yeah, that does make sense. I mean, I don't know if . You could have said, you know it if you've got the resource to to buy in the skills or whatever to do it, then that's a way of mitigating that risk. But yes, I can see. I can see both points, I can see both sides really.

H [01:07:49] I think you want you to answer about your experience. Yes. Now, could you could you, do you have the tools?

A [01:08:06] Yes, yes, I because it yeah. We're talking about and think about people using this model. You know, you'll come to it saying that this model is, you know, this is this is

our situation now. And you'll use it, to, illustrate, demonstrate what you need. So I think what (name) says is probably is quite right actually, you don't need it.

E06 [01:08:31] I actually think that's really important to remember throughout all of these questions is what we're actually trying to achieve with the model, which is demonstrating principally internally. To our internal stakeholders, the people who hold the money, that type of thing, or sign off on grant bids and that type of thing. Why we need this stuff, why this particular element represents a risk in the wider digital world. So I think that can help with some of the questions if we if we continue to think along those lines.

A [01:09:17] Yeah, yeah, I think that it. keep coming back to why we're doing this. I think is, yeah, will probably help us all actually, thank you (name).

H [01:09:26] Yeah.

[01:09:28] Can I just ask. Sorry. just I think just getting my head around the question. So out of a thousand files with file formats. Is, i,s I think when I read the second part of the question I'm thinking about, my mind kind of goes back to archives rather than a thousand files. So I'm kind of thinking, I'm sorry, I'm not quite sure how to phrase it. So, you know, how many archives would I expect to have? How many archivists would I expect to have the tools to render? And that's having the availability of tools and software to render digital material and the expertise to use them? I don't, it it kind of doesn't read right, the beginning out of a thousand files, is it not out of a thousand archives that have file formats which are ubiquitous and or open, how many would you expect to have the tools to render.? But then I may be just missing something. Sorry. That's that's just me.

A [01:10:37] These are all good questions. You know, we did this we did this internally without a very wide checking of it, wider, so it's all good feedback. If that makes more sense, I mean, that this is why I worry about the maths. Does that make any difference to the maths of the question?

H [01:10:58] So we did want to focus in this part of the model about thinking one archive and they've got a thousand files. How many? (Right. About right. Right. Okay.) So it I know it does change and apoloies that that might be a bit confusing. We can think about that. So the idea was that if think of an archive, your archive, in general how many, if they've got a thousand files, that are ubiquitous and open. Yes. Not not thinking of the UK archivist. Just thinking a thousand files. Does that make sense? I know it's a bit different conceptually to the previous ones.

E07 [01:11:39] I think I think from my perspective. So so obviously, I don't have an archive and I work at the (place) same goes for (name). So it's a slightly trickier question for me, so I could think about it abstractly or I could think about it from my experience in my previous job when I did have an archive, or I could think about it with regards to, you know, kind of a cross-section of our membership at (place). So, its different ways that I could think about this question. I don't know whether you have a preference.

E07 [01:12:19] At the moment, I have no archive and I have no tools, but I have a sense of where I might get tools and what I might buy if I was faced with this challenge.

H [01:12:29] Martine, do you have any comments on how best they should...?

M [01:12:34] My question would be, is your previous experience likely to be reasonably current, in which case, by all means, imagine yourself back there.

E07 [01:12:42] Yeah, so my previous experience, which was about a year and a half ago,.

M [01:12:47] Not long then.

E07 [01:12:48] Yeah, I could do that.

M [01:12:50] Yeah. Would you I mean, in the different ways you think about it, would you come to variety of different answers?

E07 [01:12:59] Um yeah, possibly because I think every archive is is different.

M [01:13:04] Yeah. Yeah.

E07 [01:13:08] That probably I don't know how, what (name) would do. Now (name) put in the chat.

E20 [01:13:15] Yeah, no, I've just been whimsical. I think I would just attempt that on a generic sense to I have I mean, it's been 10 years or more since I really worked in a. Institution where we were actually actively doing that. So preservation rather than my current role, though, coordinating and talking about it. So I would be dipping back into previous work and also a sense of where I perceived the community to be based on the kind of reading and writing that I'm able to engage in. So it's a different type of expertise, but hopefully still relevant.

A [01:13:51] Yeah, I think that's the point, really. I mean, we want as wide the expertise is possible. And the idea is that, you know, we get to the point where we have a good, we have good data, if you like, and all the evidence is from the experts rather than because there isn't any evidence that elsewhere. So ~~this is this is~~ this is how we get it by having people as from working in the archives or from their knowledge about the sector generally.

E20 [01:14:18] And the one question I would ask that would be a bit indulgent. But, you know, I'm I'm limiting my kind of thinking here to England, really to the U.K. perhaps, because if you had tried this question in an African context, you'd get a vary different set of results, you know? So, I mean, I'm assuming the framing of TNA means I should be thinking about English, broadly speaking, English archives with particular English public sector perhaps feel to it, but not not completely. But that's maybe that's just me making explicit one of the things I'm monitoring in my own mind.

A [01:14:56] No, that's absolutely right. And we we when we talk to, when we put this bid in, we said to the heritage fund that it would be for UK archivists.So we decided to take that approach from now, for now, because otherwise it would get you know, we'd have to consult a lot more widely, shall we say, to get opinions from across the world to make a model for different purposes for different audiences. So, yes, I think at the moment, keeping it in to a U.K. archive. Or you know, whatever your experiences is. But, you know, broadly from a U.K. archive approach would be best for now.

H [01:15:41] I just wanted to pick up some things so ~~with one of focus as well on do will you would you~~ if you just knew to file type or the file format, would you have the tool to render it even if you did open it? Forgetting about the fact it might be corrupt ~~or or an~~

~~anything~~ or anything like that. It's it's if you have these files, ~~do we~~ would you be able to, Do you think you'd be able to open it. Would you expect to have the tools. To open a word document to open a PowerPoint, maybe a JPEG.

H [01:16:26] We will have questions later on, If you open files, how many might be corrupt, how many might be inaccurate and things like that, so we have that separately later.

Question 5

H [01:16:43] Any more questions about question 4 question 5 is bit similar. So. I'm going to move to question five: so out of the thousand files with five file formats, that's are neither ubiquitous nor open, at an archive where staff have good technical skills, how many would you expect to have the tools to render? So its a bit of a different situation so, question four was the sort of easy file format. The either widely used or open? And they're sort of easy ones. These are going to be a bit more risky because they're not widely used and are proprietary, but, you have good technical skills to archive and there's a definition on the screen about what good technical skills is. So could you render it now? Bearing the mind you might have to perform file format, migration or software emulation, etc..

A [01:18:08] And obviously this, there is a definition of that as well, to define technical skills and then good technical skills. That sounds as though it's straight forward for people> not getting any questions in the chat either as far as I can see.

[01:18:36] Sorry. So. So just thinking of our question, four or five. So the numbers that I would be Thinking about is. Is on that thousand files. How many of the thousand files would I expect to to be able to render based on the previous open and ubiquitous and this one either open nor ubiquitous, so. Okay. Okay. Yeah. Okay. Thank you.

A [01:19:22] We're not assuming the technical skills are the average, No, that's I'm looking at (name) question. Are we assuming good technical skills as the average and do we pick those without good skills later, not looking ahead to avoid spoiling the suprise.(laughs) lovely.

M [01:19:44] There is also a question about full rendering.

M [01:19:50] In the chat, (name) has got a question about a full rendering or just recover some data in the document.

A [01:19:58] Right. OK. So for (name), I think the answer is we are not assuming their average. That's what we're just assuming as good technical skills. We are not assuming that every archive or every archivist has a good technical skills. (name), what would we expect to a full rendering [read's question to self]. I think, again, when you say render in my mind, we're thinking about it's good enough for the purpose of the user, which, of course, could be different depending on the aims of the user.

A [01:20:37] An archivist would probably want to render it completely, which may not always be possible.

H [01:20:44] I think here it's kind of quite key about if you have good technical skills in your archive, you will find a way to be able to render an object of this type, even if the file format was from 20 years ago. If you've got good technical skills, you might be able to find a way to render it. Eventually. Maybe not all the time.

H [01:21:24] Yes, (name) If if you don't have good technical skills and you've got a file format that is not widely used and its proprietry, you would expect them to struggle. The risk is going to be higher because they don't know what to do with this file. But if you've got good technical skills, you will either already know or you'll know where to look. And you'll be better equipped to deal with these more risky file formats.

E20 [01:22:02] Sorry, can I ask one more question, I feel like, um. Um, uh. Sort of puzzling this one out. Are we talking here about files which are held by the archive? Or are we talking about files which are kind of coming into the archive? Because it seems to me that, you know, your ingest so you're kind of accession processes might solve some of these issues. And so you will get a different calculation if it was things that are sort of at the front door or things which are already in the repository.

A [01:22:41] Well, I think it's coming in because they're going to have to worry about them only when they come in they can do something about them, assume it is

E20 [01:22:50] Things that are going to be served up to you, whether it's it all through it or some sort of document supply or whatever. Yeah. And everything from a government department or from a research or submitting them to you as a research data management process at a university.

A [01:23:07] Yeah, this is yeah, this is this is what you'll have to do deal with, you've got 1000 files given to you, but not ubiquitous or open.

E20 [01:23:13] and they've not already been processed, or quality assured or normalised in the context of the archival kind of ingest process,.

A [01:23:21] I think so. Yeah, because otherwise, you know, you can just assume that you dealt with a lot of problems, wouldn't you?

E20 [01:23:28] Yeah.

[01:23:29] ~~Yeah.~~

E13 [01:23:31] Can I add further to the point that (name) making there that. Its(name) Speaking here. Our policy in relation to that would be to, having identified that, for example, it's a highly proprietary file format, We would reject it and request the depositor to provide us with a non-proprietary version, of the information.

A [01:24:06] Ah ok, then what about when you, so there's no way you'd have something in your archive. That was not of a format that wasn't ubiquitious or open. Is that right?

E13 [01:24:18] No. What I'm I'm identifying instances where the file format might be highly, highly proprietary or or in some way, shape or form, which the archive cannot deal with but the depositor can because the depositors still has the originating software. So our policy there would be to ask for the depositor to provide an export of that information in a more, a more open format than the one that they had previously presented.

A [01:25:00] I think, go on Hannah.

H [01:25:00] If I may jump in here, I think in the model itself, there is a, there would be a question before this which would ask the user to say what proportion of their files were ubiquitous and or open and which were difficult. So it sounds like it might be the policy in your institution (name) and I know maybe at TNA that if it is a format we can't deal with, we don't take it or we ask them to convert it. And then but then we're saying that. So we will be dealing with, it might be that in your archive, all of them are ubiquitous and or open

E13 [01:25:40] But how does that go the the heart of William's question, which is where are you doing the counting? At the front door, including those that you're immediately going to reject or after it's come through because you've accepted the format.

H [01:25:57] It's saying if you have these files. If you were to accept these formats, that one ubiquitous or open, could you deal with could you still have the tools to render them? I think.

A [01:26:11] Yes, becuase if you've dealt with them, dealt with it by not accepting them that's not something we're working out a risk for here, OK? Maybe the risk is something you've got you've got to deal with, I think. Yeah.

E13 [01:26:24] OK. Right. I will. I will think that one through. Thank you so much.

A [01:26:44] OK, at TNA, we have to take things. If the government department has deemed it appropriate for, its record and te appropriate for permanent preservation, then we will take it, no matter what format its in and worry about how to look after it later. Usually I ask (name).

H [01:27:07] OK.

A [01:27:12] Next one? Yeah?.

Question 6

H [01:27:18] Uh oh, let me change the slide, there we go. Right. So some more definitions here. So we're talking about out of a thousand born-digital files, how many would you expect to have the content metadata that meets an archives requirement? So you should all have a definition of content metadata. And of born-digital records or born-digital files, so take a minute to read the content metadata. If you haven't already. Again, we know this will be subjective because to meet an archives requirements, archives might have different requirements.

E07 [01:28:29] You know, this is an interesting one. And my my instant kind of thought is that most archives won't really know their requirements, for born-digital content metadata, I don't if that's a fair comment. So I kind of tend to think this more in the, in terms of the requirements of an end user, like the requirements that will enable someone to reuse reuse that content in the future. I don't think that really helps around framing the question, but it's just an observation.

A [01:29:12] I suppose if people don't know what there minimum metedata requirements are. Martine, would that mean that the risk would be higher? Because they just don't know what they need. Well, I haven't decided what they need. Not sure how that impacts.

M [01:29:39] Content metadata, that meets you're archive requirements. Presumably so that you could then process and.

A [01:29:45] Yeah, you've got intellectual control and you've got the ability to technically preserve them. If you don't know that obviously the, the risk will have to go up, won't it. Because you'll just take what you are given without asking for anything more.

M [01:30:06] If you don't have the metadata. What effect would that have on? Would you not be able to render it? Would you not know how to approach it. So provenance information. Is it really genuine? Has it been altered by preservation, etc..?

H [01:30:25] And so about the state of the process, is this pre or post ingest? I think I'd say this is thinking just I would think of ingest. How many do you have the content metadata that meets the requirements, be that because it came from the depositor directly. Or maybe you filled it in yourself.

A [01:30:48] And I think we're assuming from what from (names) question is that it is actually correct as well. You've got it in the archive. You've got metadata. It is accurate.

H [01:31:05] Yeah, I guess it'll be difficult for us to predict how much of our content metadata is inaccurate. Just have we got something in those fields that says, well, whatever is your minimum standards maybe the title might be depositors' name.

[01:31:26] So does all this, So does all this information have to be within, you know, our content management system? So for us, some of it will be but some of it will also be within, you know, like the accompanying donor file, which is still, you know, doesn't and isn't currently attached to the accompanying record for that item within our system.

A [01:31:53] Yeah I think we don't mind where you're keeping the information. It's just thatyou've got information.

[01:31:57] We've got information. Great. OK.

A [01:31:59] Yeah.

[01:32:00] Thank you.

H [01:32:00] We do have a question later about joining the things together, and about do you have good information management. But yeah. For this case just do you have it.

[01:32:07] Perfect. Thank you.

Question 7 and 8

H [01:32:22] I might move on to question seven and eight, because they're very similar. And let me set the slides. So it's the same question, but for digitised files and surrogate files. So my understanding of this is, well, actually, Alex, as a TNA expert, and they still want to explain the difference?

A [01:32:46] Because digitised is one of those words where it's got general meaning. And we wanted to be, We've got a very specific meaning in this contex for digitised in that is records that were originally analogue, so photographs or paper. They've been digitised

and those digital versions are the official records that you've taken in to your archive. So you haven't taken the photographs. You haven't taken in the files. So digitised is slightly, it's a slightly unusual meaning, I think, or for this context. But that's what we want you to think of. If you don't do this, then I think it's just a question of imagining. If you, if you had files of this nature where you couldn't go back to the original because you didn't have it, how would you feel abou, htow many would you expect to have that content metadata that meets the archives requirements? You know, would you expect it to be more more stringent, neglected or or not? Because you're fine with, ~~digitising~~ the digitised version your you're trusting is fine.

H [01:34:01] And then surrogate is the same situation but you hold the analogue.

A [01:34:10] Yes surrogatea is what normal archives have in that it's just a copy of what you've got in your archive. And it's not considered the record, the official record. You know because you'll always have the paper or the analogue original to go back to. I think that's generally, yeah. I believe that's more common use of the term generally, I think in archive as well. But again, I might be wrong. Limited experience outside of TNA, I am afraid. Because I've been there since the last century.

E20 [01:34:48] Break it to you, Alex, that was the last millennium.

A [01:34:51] Oh, god, thank you, (name). That's so helpful. Yeah, no, I'm feeling great now.

E20 [01:34:57] We aim to please.

Question 9 and 10

H [01:35:08] I'm going to take silence as assuming there is no more questions about six, seven and eight. So they've all were very, very similar question. But one's born digital, one's digitised and one's surrogate. OK, next questions nine and ten. So we're moving again here and instead of files, we're going back to a thousand U.K archivists. So question nine: out of a thousand U.K. archivists, how many would you expect to say that their catalogue management system met the needs of their organisation? And then I will read question ten because it's similar: out of a thousand U.K. archivists, how many would you expect to say that digital asset management system met the needs of that organisation?

E07 [01:35:57] It might be useful to have a definition of digital asset management system and whether that also covers digital preservation systems.

H [01:36:10] Yes, thats. Probably a very good point, actually. And.

E02 [01:36:15] And also current versus future needs so.

A [01:36:21] So, again, I think I was working on, are we were all we always working on the premise that it was current?

H [01:36:30] Yes.

A [01:36:31] That's all you can do really is if you could all you can do is, is is work with what you what you need now. I know we're always trying to look into the future as digital

archivists, but we can only ever look so far into the future. And focussing on keeping things alive for now so that they survive into the future is the best we can do most of the time. So.

E07 [01:36:57] I suppose another thing to say on this one is that some archivists don't have a digital asset management system or a digital preservation system. And how you would answer that question, given that? It depends what you define. I Really, Yeah. It does depend on the definition of that term I guess. 04: Yeah I agree.

D [01:37:23] I can think here where it is, say there is no system. Then by definition, it's probably not meeting the needs of their organisation. So, you know, we're we're. Partly appreciating the fact that, yes, some people are out there with with systems that aren't of purpose.

H [01:37:42] So I guess the system, so it could be some advance sophisticated software or technology. But then a catalogue management system could just be a paper list that that is well maintained. And, you know, it does the job fine. And so it doesn't need to be very sophisticated. I don't know if a digital archivist wants to expand on a digital asset management system and how it's different to a customer management system.

E07 [01:38:12] I think that catalogue management system is managing the metadata, really the descriptive metadata about the collections. It's my understanding and additional asset management system is managing the actual digital objects.

E02 [01:38:26] I would agree that we have it split.

[01:38:30] Yeah, same here.

A [01:38:34] Ok that's good, but we should probably have a definition of that I think written, somewhere that would be helpful. Thank you all.

[01:38:46] ~~All right.~~

M [01:38:48] There is a question from (name), I assue this is a yes no answer rather means requirement 80 percent of the time.

H [01:38:54] I think I've just come in there. So, yeah, they would say yes, no. But you're going to answer out of a thousand, how many would say yes.

A [01:39:06] Right.

H [01:39:09] Um we'll put up, we'll find a digital asset management definition, and I'll share that with you. During the lunch break. Just to be clear on that.

E02 [01:39:23] Yeah, I think some people might split it out between repositories, too, so you might have three layers.

H [01:39:28] OK.

E07 [01:39:31] And some people will, who don't have a system, have got processes and procedures instead of a system. So they might have storage well you'd hope they'd have storage, but they have a ~~system,~~ a series of processes like they run their own checksums

over that storage or they they have a spreadsheet they record that metadata in. So it's not a system, but it is still a kind of a it's still doing something that might well meet the needs.

A [01:40:00] It's still well managed isn't it, or could be well managed. Yeah. So might want more broadly than I suppose. Yeah. As a system that could be make up made up as a step of workflow steps. Yeah.

H [01:40:16] Yeah. It does not need to mean some sophisticated expensive application managing all your files. It can be basic, it just something organised and making sure it meets your needs.

Question 11

H [01:40:28] Yeah. Yeah. Okay. OK, I'm going to. I'm going to move on to the next slide. So again some definitions here for you, question eleven: out of the thousand files with insufficient content metadata, at an archive where there is sufficient information management, how many files would you expect to be able to identify, i.e., knowing what they are and where they are from? So we're now talking about situation we don't have good, we don't have the content metadata we need fully, but we do have good information management and there's some definitions up there, so this one's a bit more, specific.

H [01:41:34] So some of the content metadata might be missing, but you still know where the file is. You can still know enough about it to be able to identify it.

E13 [01:41:49] Hi, it's (name) here. Can I use what we're learning in question eleven just to clarify my thinking about six, seven and eight? Well, I had previously thought I was clear. But now I believe I'm not. In six, seven and eight, is the implication within the question that the content metadata is embedded in the file? That certainly could be the case since content metadata refers to author and title, which a lot of, for example, word processing software sticks in the file.

D [01:42:40] I think provided you would know how to get that out. And then for this sort of thing, then the information management and so on would be enough to do that. That's probably okay. Yeah.

E13 [01:42:52] Well, I'm I'm thinking back to my reading of six, seven and eight because I was allowing the content metadata to be on a separate piece of paper. Yeah. Is that acceptable in six, seven and eight?

D [01:43:11] I think so. Yeah.

E13 [01:43:12] But in 11, what we're talking about is a more formal information organisation for identifying the for example, the provenance of the of the object.

H [01:43:39] Yes, I think so. In question, eleven, we are talking, a step further from, you might have the information, but do you have it, sufficient information management means you'll have it in, you'll know where it is to go and find it. It won't be a piece of paper in a box somewhere. It will all be, you know, enable you to easily find everything about that file. And with that, you need sufficient information management to identify it.

E13 [01:44:12] Does this go as far as discovery?

A [01:44:17] We're keeping away from access and discovery actually, we decided that would've been a whole other section of the of the model that we wanted to that we'd want to do in depth. So I think the distinction and please David adn Hannah correct me if I'm wrong, distinction between we have to was, the content metadata was around with understanding the record and its meaning, and information management is around knowing where you got it. And it's kind of custody and a bit more around that. Yeah.

D [01:44:53] So I guess, yeah. If you if you had opted to just leave the metadata and just use what's associated actually in the file, then as long as you've got processes and tools that would allow you to do that then that's fine as part of the information management.

E13 [01:45:12] Excellent. In order just to test that I've got this right. If I go back to question six where it says that meet an archives requirements.

D[01:45:23] Yes. So if that's the decision that the archive has taken, the just that the internal content. Yes. File metadata is sufficient. Then.

E13 [01:45:32] So I'm. I am suggesting that I can read that as meets and archives initial requirements.

D [01:45:45] Yes, I suppose so, yes. And then did the information management processes and so on here, if you've got those good processes and that's fine if if you've sort of made that decision, we'll just rely on the content metadata. But then you don't actually have all the metadata associated with the file. You don't actually have the processes and stuff to remember how to do that. So when you actually try and find a specific file, you can't do it, then it's, you know, that's breaking down. So. Yeah.

E13 [01:46:13] OK. Thank you.

Question 12 and 13

H [01:46:25] I'm going to. I'm gonna move on then. If no one else got anything else, particularly on this one. Question 12 and 13. So we're moving slightly more technical now. Out of a thousand files at an archive where staff have good technical skills, how many files do expect to have sufficient technical metadata and the same question again but when staff have poor technical skills, how many files would you expect to have sufficient technical metadata? So take a moment to familiarise yourself with the definitions of technical metadata and good technical skills.

H [01:47:13] And an example of how you would get the sufficient technical metadata, you might run a droid report or use various other validation tools to understand that. And you need to understand the technical metadata later to know whether or not it is a full vendor.

E07 [01:47:43] Really, my kind of query on this question is just about the words sufficient. And so what we always find and then what we tend to say to people that we give advice to is you need to understand kind of your users and your future use case in order to know what's sufficient and what isn't and so I just find this one quite kind of tricky to start to think about an answer to in an abstract way.

H [01:48:27] So I think. When, I had discussions with people here at TNA about this, we agreed that it's always difficult to know what a full vendor will ultimately be. So another way of perhaps thinking about this question is, in terms of sufficient technical metadata, had

you run a droid report or done some sort of validation tool to know basic characteristics that will be important for the file. This might be difficult to define. I don't know if any digital preservation expert, David or David, wants to jump in and explain what an example might be of that idea of sufficient technical data.

D [01:49:06] Yeah, I think. It is obviously difficult to be to be certain. But the fact the very fact you are thinking about what the range of possible uses are means, you are more likely to get a fuller set of technical metadata, so always it is always a bit chicken and egg for this but yeah. If you've got good technical skills, you'll thinking about different ways. It might be access. Different questions people might ask of the files. And things like that say, yes, some of your the word star files that you analyse you know, someone was just doing like sentiment analysis, then just being able to get the plain text out is fine. If you're actually wanting to understand more about the process, knowing the formatting, knowing that, you know, one scene was expected to fit on a one page or whatever it was, then things like that become more important. So, yeah, having thinking about the range of ways that people might want to access it, what they need to do later is all going towards the sufficiency of the technical metadata data. Yeah. I mean, this is a problem. We can never be quite sure how much we do need to be trying to extract at the moment anyway.

H [01:50:28] And it is about being able to extract it. I think if you have good technical skills, you can extract it. Maybe if you have poor technical skills, maybe you can extract it in a few cases.

E04 [01:50:44] It's also knowing about which tools are out there and which tools might apply for a specific file types.

H [01:50:53] Yes. Yes. OK. I'm gonna click on.

E13 [01:51:03] Before you do before you do, its (name) that again. I'm I'm I'm still trying to get my head round this from the definition and then the description of technical metadata. I, I'd always kind of thought about this is the technical metadata being the metadata which is intrinsic to the digital object. Now, the the the description that we've just heard, I can only make sense of in the context of digital object having been migrated and in some way, lost its original technical metadata, because I wouldn't expect to be editing technical metadata. In fact, the the whole point of it is that you don't edit it.

H [01:52:02] That is true. But not we're not editing it.

E04 [01:52:07] It's about collecting, isnt it, essentially. Knowing what you want, what the technical information you want from a particular file type is and which tools you use and what you're expecting from that, a particular file.

D [01:52:21] Yeah, I mean, it's kind of what you would want to hand have to hand if you were doing a migration or an emulation in order to be able to check that those things had had then represented the file well. So from an image for your probably, going to want to have information on the pixel dimensions and things like that. So, yeah, you'd expect you probably would be able to go back and do that, but it may be easier to have extracted that once during your ingest process and then you don't have to actually read the whole file and things like that. So yes, they're intrinsic to the file, but you will often want to have them stored separately to avoid having to access the files so much depending what where you're storing them. You know, there can be costs to then actually accessing that original

file. So ways of being able to compare it and have it to hand in an easier way. But yes, it is stuff essentially that is intrinsic to the file. Is that information about the file format itself.

E13 [01:53:23] So so this does rather begs the question how kind of file have insufficient technical metadata? Because a file has got the technical metadata that the file's got.

H [01:53:37] So the difference here is would would you be able to know that information? And have you looked at that, inspected that file and worked out there are five different audios. So that the file will have it, but do you have the skills with the archive with good technical skills or poor technical skills to know how it should be rendered?

E04 [01:54:02] Hannah's descripton is pretty good. Because you could have a video file that has one audio track or another audio, another video file may have up to eight audio tracks, possibly. So if you were thinking about doing the digitisation from one format to another, if the format you're putting it tocant support eight audio tracks you may will lose data. So it's having this information at hand from the original, so you can assess what you want to do with the future or if there's any digital change that's to be noted.

E13 [01:54:35] Well, yes. I, I think I understand all of that. But my point is, as far as any given file out of these thousand files is concerned, it's only got the met the technical metadata that it's got.

[01:54:54] Yes.

E06 [01:54:55] Question is, is perhaps worded slightly misleadingly since the question at the moment is reading. So does the technical metadata exist or doesn't it? Well, actually, it always does exist. So the better way to think of the question is not not how many files would you expect to have sufficient technical metadata? It's how many files would you expect to have extracted the technical metadata.

D [01:55:27] Or for for which you, yeah, you either have it or can can get it. Yeah. I suppose that's the key thing about thinking about where people have good technical skills as well.

H [01:55:40] We did have a bit of difficulty wording this question, so.

D [01:55:44] Yeah. If you got poor technical skills so poor you don't even know how to run droid then probably don't really know what the file formats are.

E13 [01:55:52] Yeah.

D [01:55:54] Okay. So yeah, maybe we could have got got the wording a bit better. But I take your point. Yes, it is intrinsically there, but unless you can actually have those skills to access it, it's no use to anyone.

E13 [01:56:05] Right. OK. Thank you.

H [01:56:14] I don't think we need to assume there is a system that we have that we put the that that extracted technical metadata in. But I think we need to question about can you extract a technical metadata if you want? Because you've got the good skills you can maybe if you've got better skills, you are more likely to be able to extract it.

H [01:56:43] Shout if you're not. Otherwise, I'm gonna move on.

Question 14 and 15

H [01:56:54] Getting a bit more specific here, moving into storage life and specific storage media. So question 14: out of a thousand hard drive disks, kept in a monitored commercial environment, how many drives would you expect to fail in any 12 month period? And question fifteen is the same but how many would you expect to fail within their first 12 months of use?

E13 [01:57:24] Sorry to be bombarding you with so many questions. It's it's use of the word fail. Do you mean that the drive fails or do you mean that there's a loss of data?

E04 [01:57:39] You're talking about drive faiure aren't you on this one?

D [01:57:42] the drive fails. So that might lead to a loss of data. If it if that was the only copy of it. But it's we're talking you know, we're talking about a server farm or [E13: a monitored commercial environment.] Yeah. Let's say how many you know, if you just got a whole bunch of drives running. How many of those will actually fail in that in that 12 months? And we're not talking you know, when we say commercial environment, we're not talking about an external hard drive that's been put on the shelf out the way.

E13 [01:58:13] But but we are agreeing that there is no loss of data.

D [01:58:18] Not necessarily any loss of data, no.

H [01:58:22] Yeah. And a commercial, maybe of commercial is an unnessary word, I've sort of wanted to emphasise we want to talk about people who know what they're doing, who are looking after hard drive disks, who are monitoring it, not sitting on a shelf or sitting, getting dust in someone's garage.

A [01:58:49] Yes. I suppose it could just sit in a monitored environment, couldn't it? Yeah. Yeah. We're not talking about a storage facility run by Google or Amazon. We're talking about an archive with there own storage facility, their own archive storage.

H [01:59:06] Yeah. But a step above me doing it at home. OK. Any other questions about 14 and 15?

Question 16 and 17

H [01:59:26] Right, I am going to tick on, because the next question's a bit similar as well. 16 and 17, so: out of a thousand, NAND solid state drives kept in a monitored commercial environment, how many drives would you expect to fail in any 12 month period? So, again, commercial doesn't necessarily mean commercial, but monitored not my garage. And then the question, 17: out of a thousand an NAND solid state drives kept in a monitored commercial environment, how many drives do you expect to experience a persistent read error within any twelve month period? So not as hard as the drive itself failing. But now we're talking more about a particular error.

H [02:00:28] And these might be hard to answer if some of you aren't too familiar with particular storage medium, and that is fine. There'll be some questions that you'll feel more familiar with and more expert in and others that you may not.

H [02:00:49] Any questions before we move on?

H [02:00:57] ~~OK. Oh, sorry.~~

H [02:01:04] So, (name). If you have absolutely no idea, I don't know, Martine should she do, should you pass the question or just put a very wide guess?

M [02:01:17] So the uncertainty bounds, the five percent in the ninety five are that to tell us about the uncertainty you think is in the system or the uncertainty that you have. When ee have a discussion tomorrow if you wish to you can say, ah yes, I put that, because nobody will know unless you do, and I was really uncertain about that. So put something and then we can address your questions when we get there. And what we will actually prioritise is what you do in the second round. So the first roundis a basis for discussion and we want it to be as accurate as you can, but if at the moment you feel like you don't have enough information about that, hopefully after discussion tomorrow you'll have a bit more information and you'll be able to adjust. This is not your only go at this. So if you could put something that would be helpful.

A [02:02:16] It's also a question from (name).

H [02:02:19] So I will have, if I understand it correctly, I think. I think you're asking about even if you have poor technical skills, you might be able to get some technical metadata. So how is that?

E05 [02:02:32] No, I think I think they're struggling. I mean, I struggle with numbers in general anyway, so I might just be really missing the point. But I think what I'm saying is, so if I were to say someone had poor technical skills and I said, okay, well, the lowest I think they could possibly be, I'd expect to have a thousand is zero. Right. But then if I expect anything above a thousand for any of the others, well, then why couldn't they get fully to a thousand? Because if they found a tool to extract technical metadata, for example, well, then why would they apply to all thousand files? So then would the answer always be a thousand if anyone can do anything above zero? But then I think maybe just confusing myself.

H [02:03:07] I think. Think of a thousand random files.

E05 [02:03:13] Ok that maybe needs multiple tools and multiple types of skills or something.

H [02:03:19] Yes. So you have your thousand files, might, won't all necessarily be ones that could be done by Droid.

D [02:03:28] Well, they are probably, you know, you've got a mix of image files and video files, so you might know how to use image info to extract some stuff out of an image file, but you don't know how to use some of the tools for video and stuff like that.

E05 [02:03:39] Okay. okay cool, thanks.

Question 18 and 19

H [02:03:46] I'm going to move to question 18 and 19. So we, again, still thinking of storage and it's a slightly different concept now. So we've tried to define a generic category of storage media because there are lots of different ones used in archives or in general. So we have a type A that is less stable and we're saying less stable in terms of unexpected lifespan or is highly susceptible to damage, or it needs very specific conditions, sensitive to changes, prone to errors, et cetera. So big bucket. And we know everything's on a scale. But in general, a sort of less stable when we have some examples there vs. a Type B, which will be more stable. So maybe what we expect it to live longer, not be as susceptible to damage, but more robust could perhaps live alright on a shelf for a year or two.. whereas type A wouldn't. I mean, some examples there as well. So we know this is a bit subjective. So you might want to discuss what, get in your mind exactly what might in A and B, and we have got some examples. And so a couple of, few of these questions talk about that type A and B so question 18: is out of a thousand media of type A, which is the less stable type, how many would you expect to reach the end of their life within 12 months? and question 19 is identical for type B. So the end of the life we'reing talk about fail. We talking about the storage device fail. As we've been doing before. And we would expect that to be a range because, you know, depending on what type of type A device maybe USP flash drive, you know, has a higher chance or a lower chance of surviving than an SD drive. So we know there probably will be variation, uncertainty particularly by categorising a type A or a type B device. We shouldn't be doing any Googling. No, that that's a good point. So please try not to Google you use Google or things to help you with this.

H [02:06:05] So, about (name) question about type A. This is in reality you might want to move your type A's to type B, as you might be an archive that as soon as anyone gives you a less stable storage device, you'll move it. You don't want to own a type A. But then that's the sort of separate question this is, if you did have type A, even if your policy is no, we're never gonna take that kind because we know it's not stable, how many would you expect to reach the end of their life? So I imagine lots of you probably do, in your archives themselves have stable, more stable media. But if this would give people an argument of, well, look, if I if I didn't have it on these more stable media I think a lot are going to not survive the year.

H [02:07:00] Any questions about.

[02:07:03] CanI just checked, sorry, so just thinking about end of life. Couldn't that also mean that? And so thinking about LTO tapes, that that generation is no longer supported?There may be no longer drives available for that generation or something like that? or is it purely that you get it off the shelf, you mount it and it fails?

H [02:07:31] So there's a seperate questions later about equipment. So we're not talking about you don't have the drive or you don't have the software. This is very much assuming just the physical drive itself. When you take off the shelf, it fails.

[02:07:48] OK, thank you.

[02:07:48] Yeah. Yeah.

E20 [02:07:54] So I am kind of struggling with this a little bit for a moment, if I may. What what role do warrantees play in this? I'm just thinking that certain media types have warrantees you know or they have claims associated with them. But obviously the warranties aren't necessarily reliable. They're always sort of modelled in some way.

A [02:08:18] I don't think we're worried about warranties. Because this is about losing data, about something actually failing. So even the best warranty in the world isn't going to recover your data. It's about the stability of the media, I think.

H [02:08:39] That this is why everyone might have different experiences that they can offer here. So some people might know I use this type of storage media and even though I've seen publications that say it can survive twenty five years, I know that's ridiculous and only tends to last five. Sowe know there'll be a range here and everyone will have their different views of how long things will actually live. And so we're not necessarily going to take what an LTO tapes says on its disk, we want your opinion, be it the same as what is published and what the providers say or be it from your experience, actually I do think it lives longer than that or actually there's a higher chance it will fail within the first 12 months or in any twelve months.

E20 [02:09:27] I mean, that's quite important there's a lot of smoke and mirrors really in the fields of storage media. There are some outright lies and there are some fiscations, half truths that are stated very frequently.

H [02:09:50] We'll talk about, some question about whether it's a brand new media or any media. So I did we intend this question to be you don't know how old is. But don't think at the end of its life. And I don't know if, david or Alex, you think that's the right way we should be thinking about this?

A [02:10:20] Mhh, I don't know because you often do know the age, don't you? But we know exactly the age of the tape or something. I know we've got lots and lots of hard drives, and I don't know if we know exactly when, how old each one is.

H [02:10:36] Yeah, so I think it's probably got to be we know this to be difficult and they'll be maybe big range just on average, yeah, some will be some you have to be younger, some will be older. What were the chances of it failing next year? Within within the year.

H [02:10:57] And again, this also doesn't take into account the conditions in which they're stored. So at one point, we were going to include that but to simplify it, we haven't. So, you know, some if they're kept in it's totally optimal conditions you think they'll live forever, but you might know actually these are really susceptible to temperature changes. And this question doesn't say it's in a monitored commercial environment, so maybe more would fail sooner..

E04 [02:11:29] I think that is a major thing maybe to think about though. It tends to be, certainly with magnetic tapes, and the like is humidity is the major factor possibly more than temperature. In many cases I can maybe, it can quite drastically affect longevity of an item and half it.

H [02:11:59] Maybe we should think it. Would it help if we say, think about this from your experience of how an archive would store the media?

E04 [02:12:12] I think maybe we might have to maybe it's us, yeah, it's, not every archive is going to have optimal conditions for storage is the thing, isn't it? But it is a major factor in the longevity of materials. Maybe we do need something in there. Maybe as maybe possibly as a separate question or an add on one of these.

H [02:12:43] Mm hmm. So I think we were initially hoping that in you, in your answering the five percent, 95 percent, you can reflect maybe that range of, Well, if it was kept in, not very good conditions, that might be your five percent, and if it was kept in fabulous, it's from nine five or allow a range. But if you say that's very difficult, we can tailor it. Do other people thing to be quite difficult, there's would be lot of fluctuation depending on the conditions that might just mean the range is a thousand? Maybe the range is a thousand.

A [02:13:27] Mhh. Would that be a useful way of thinking about it for the maths of it, Martine, if people are less sure, they use the extreme percentiles?

M [02:13:44] Yes, I think that's what we want to know, because there will be a range there won't be an answer for this. It will be if we had data on this we'd say, well, on average, they last seven years, they can fail after two, and some of them are still going after 15. So we're trying to replicate the data that we don't have on that.

A [02:14:03] Right. OK.

Question 20 and 21

H [02:14:12] I'm. Going to move on to question 20 and 21. Because they are a bit similar. So, again, we're talking about storage, type A and B. But here Iwe are was saying, imagine there was a flood at your storage location and the archive has inadequate mitigations, out of a thousand media of type A and then of type B, how many would you expect to be destroyed?

H [02:14:44] So, again, so we haven't got any mitigations. So imagine, you know, it get it gets flooded and maybe you imagine that type of media is submerged in water. How many would just instantly all be destroyed? And be worthless and not have anything? Or how many actually might still remain intact and be usable afterwards? Does that make sense to everyone?

H [02:15:30] OK. I'm going to assume that hopefully that flood question people are understanding. So it is. Yes, saying there was a flood. You don't have mitigations. You know, how many of your disks of type A and storage of type B would just be destroyed, not usable ever again? So, yeah, you might say, oh, well, they're all gone. Or actually there's some. So I'm going to I'm going to move on from that one.

Question 22 and 23

H [02:15:56] Question 22 and 23. Again, some definitions to help. Out of a thousand files, all stored in a storage media of Type A, which is the less stable, at an archive which has staff have good technical skills, how many files would you expect the bitstream to be inaccessible due to obsolescence? So this is talking more about the equipment, the hardware, the software that you need to be able to access the data of that file. So we're saying that ~~you've~~ you've got that media, and you have good skills, but how many could you still not be able to get the data off? If it's a type of type A? And then question twenty three is identical. Still type A but we're saying you have poor technical skills. So it's the less stable media, you don't have that good technical skills. How many can you not access because you don't have the supporting software or equipment?

H [02:17:16] Any questions or clarifications about this concept of accessing the bitstream or not being able to access the bitstream due to obsolescence?

E10 [02:17:26] Hi, yes, it's (name). Here, just. Just in terms of good and bad or poor technical skills, even if you've got no technical skills, do would it not be reasonable to assume that you would just ask somebody, anybody out there, jisc mail list, or TNA ~~or or~~ or somebody else? So how how is the technical skills, a measure of of the files being available or not available or accessible or inaccessible?

H [02:17:59] So I think this is a bit similar to our previous thoughts about outsourcing of asking for help. And you're right, this is about asking the wider community and then maybe they'd be able to help. I'm tempted to say, but Alex or David, if you disagree please shout, that we should be thinking about within your organisation and without asking for help, because if you did ask for help, then then maybe that's sort of you have had access to good technical skills.

E10 [02:18:32] So asking for help within your organisation then?

H [02:18:36] So thinking about this more than archive level, So if your archive in general has good technical skills, then within your archive, then that could be anyone within your archive, It doesn't necessarily need to be you, but if we keep this of within you archive focussed.

D [02:18:53] If you've got the connections and so on with I.T. and they've got those skills, then that's fine, I think. Yeah. Obviously, yes, there is a lot of discussion that goes on, but at the end of the day, you probably can't. You know, asking and getting assistance from that wider group is less likely to be effective than actually having the skills yourself. And yes. If you know exactly what you've got and so on, you might be able to find a contractor who can actually pull something back off a particular storage medium. But just asking around on archives NRA or whatever, you might get some pointers in the right direction, but it's going to take you a lot longer. So on.

E10 [02:19:49] Yeah, ok , thank you

A [02:19:52] I think to ask people to help me with it at TNA, because there are some very skilled people like (names), you know, things have been got done that I couldn't actually do it myself, even if I had sat there and taken me through it. It still wouldn't have happened but because they were in the archive, so that kind of helped, but if they were trying to do that over Zoom, I don't think that would have worked.

Question 24 and 25

H [02:20:24] Question 24 and 25 are identical. But it's not talking about type B.

E13 [02:20:28] I think before we move on. Yeah. If I can just interject there. The I think I think the issue is is more about whether or not one recognises that even though the individual themselves doesn't have the technical skills, that they are aware that that's not the end of the line as regards accessing the bits. If if an archivist with no technical skills, but with the knowledge that a dead drive can be resurrected, then that is quite important. If they don't have that knowledge that the dead drive is skipped, then then that's the alternative. So it's not about how readily accessible the recovery path is, it's more about the individual, the professional knowing that a recovery path exists.

E14 [02:21:38] Yes. I mean, some or all these things, it's sometimes about knowing the questions to ask and where you, where can you go from here?

E13 [02:21:45] So that's the distinction that I would have.

H [02:21:50] I think that's helpful, I'm thinking about it. Thanks (name), yeah. So, yeah, you might not necessarily know exactly how to, how to access the bit stram for that particular file but you know that it can be accessed and if you ask David, he'll know a way how. And there's a question twenty four and twenty five, all identical apart from it's talking in terms of type B, rather then type A. So it sort of is the same question four times, but good and poor technical skills. And type A and type B. So because of the way we define type A and type B, Type A was a more volatile, less robust media compared to type B.

Question 26 and 27

H [02:22:48] I'm gonna move on, but speak quickly, if not. OK, question 26: Out of a thousand UK archivist's how many would you expect to say that they had at least some knowledge or skill to be able to generate checksum of a digital file? So this doesn't necessarily mean they can do it, but they know would know how to.

H [02:23:29] Any questions on 26?

H [02:23:36] And question 27 on the same screen: out of a thousand UK Archivist's, how many would you expect to say that their I.T. provider supports the requirements of their organisations archival activities to a large extent or a very great extent? So some kind of support, they would say. they have, you know, I.T. support for their archival activities.

A [02:24:12] Something we've talked about at the previous workshops as well for tthose of you who were at them, about how your hands are tied sometimes by the limitations on what I.T. will do for you when you're part of a larger organisation and archiving isn't the be all end all of your organisation's remit.

Question 28 and 29

H [02:24:40] OK, if there's no questions there that I'm going assume they are alright and move on. So question 28 and 29. So, question 28 is: out of a thousand UK organisations which experienced a cyber security breach or attack, how many would you expect to have experienced a virus it be experienced by viruses, spyware or malware, including ransomware attacks? So this is out of a thousand organisations that have had a cyber attack, out of those, how many do you think would have been would have experienced virusess, spyware, malware, or ransomeware? Again, if this, if you have no idea, don't worry. Some of these questions will resonate more with some people than others.

H [02:25:38] And question 29 out of the thousand global data breaches, how many would you expect to be due to system glitches? So, again, thinking of cybersecurity incidents, how many were system glitches?

E20 [02:25:54] Can I come in there and just ask about what we mean by glitches, so I'm guessing, to question 28 relates to some sort of deliberate, determined can have created, You've had to use the word attack there. Yeah. Where as question 29 means there's been a data breach but its happen almost by coincidence, by perhaps per design or poor implementation of a system, but without any malice or a forethought.

H [02:26:26] I think that's a very good way of that distinction there. Thank you, (name). Yeah, that's how I would see it. So the first one is a malicious attack and the second one is, sort of an accident, or coincidence, maybe it could have been stopped if we'd been a bit more diligent, but I wouldn't call it a deliberate attack.

E13 [02:26:50] If I can come in there. So as far as the system glitches concerned you're including misconfigurations, and, for example, leaving public Amazon buckets and the like.

A [02:27:15] What did you say about the Amazon bucket thing?

E13 [02:27:21] Leaving publicly open an Amazon dead bucket so that's is publically...

A [02:27:28] So setting up the system wrong.

D [02:27:32] Essentially, yeah. I think that's a reasonable example, I think. I think it used to be the default as well, that it was it was open by default. So unless you actually or some of the things we obviously we've seen on Zoom recently, you know, the defaults on on Zoom were quite open to start with. And whilst it was only being used a relatively rarely, that was fine. But as soon as it got a whole load of attention, then you got the Zoom bombers who were just trying to look for ways of finding ar random meeting and interrupting it. So yes, there's those things can be kind of out there latent and unless posed accidentally or yes, sometimes someone will go looking for it. But..

E13 [02:28:10] So if you go back,.

D [02:28:11] It doesn't have that element of the additional virus and spyware or malware being being put onto it. It's just someone configured something badly.

D [02:28:20] And someone has discovered that and done something [E13:Opportunistic.] yes. Yes.

E13 [02:28:25] But if we go back to 28, a lot of malware is just out there looking for vulnerabilities. It's not targeting a specific site at the outset. It's looking for open sites or vulnerable sites. So you could say you could argue that a malware attack is as a result of a glitch because the vulnerability hasn't hasn't been patched.

D [02:29:00] True, there's a little bit of overlap, but I think it is it it's that malice of forethought. The malware is out, is there looking for those, but a system glitch it might be something temporarily changed or something. So, yeah, that there's always there's some overlap between those things, but it's the level of of organisation and deliberate exploitation, I suppose.

E13 [02:29:28] I think I remain unconvinced. Okay. Thank you.

Question 30

H [02:29:41] If there are no more questions, I will move on to 30. OK. So question 30. Out of a thousand UK Archivist's, how many ~~would you~~ would say that the digital collections were fixity checked at regular intervals and by fixity checks? I think most people for most people that's checking the checksum still match. Does that question make sense? Anyone need any clarity.

E07 [02:30:15] The word regular here is just is just regular, not frequent, so regular could be once every five years, which might not be acceptable, but it's still being fixity checked regularly because its once every five years. or, or are we thinking about unacceptable frequency as well.

A [02:30:39] I think we were implying a regular frequency, don't you (name)? Because it was about, you know, if its worth doing fixity checking its worth doing it regularly to ensure that.

D [02:30:54] There is always that issue that you could say, yes, you're a regular church attender because you go twice a year at Christmas and Easter but you're not a frequent... Yeah. so there's probably a bit you know, it's something that's being done again, yeah, regularly at a frequency that is actually likely to be to be useful as well, I suppose. It's not that you'll be doing it once every 10 years, by which time if you discover some an issue, then actually it's too late anyway.

H [02:31:28] I think a regular would sort of maybe planned as well. It is not sort of a spontaneous you have decided you are going to check your collection, be it that every week or every year.

E13 [02:31:45] Can I come in again here? It's a bit techy, I appreciate, but in question 30, do you draw a distinction between explicit and implicit fixity checking. What I'm thinking of is the file system such as ZFS implicitly carry out continuous fixity checking.

D [02:32:16] Yeah, it's an interesting one, because I mean, we rely to a large extent on the fact that our tapes do some verification when anything is read off them. And if that highlighted anything, then we would actually check the checksums. So, yes, I think that potentially counts. I suppose it also depends how much you actually for that kind of thing it depends are you actually checking the logs and seeing if lots of things are being fixed? And that's probably suggesting that you've got other issues going on. So you might want to take other actions if you're just completely relying on that automated process, you perhaps haven't got enough real visibility of what's going on. And it's the same with maybe with some of the cloud things they offer. All sorts of guarantees and so on. But it's actually having that visibility that the checksums are not changing over time. So you might run some of those actual checksum checks less regularly and on the basis of a trigger from from something else like that. But purely relying on those is probably a bit dubious.

E13 [02:33:29] Well. Question 30 doesn't really enquire as to any kind of management actions or activity one way or the other. It just asks whether fixity of files is checked at regular intervals. And if you happen to know that your files are all stored on ZFS, then the answer is quite clearly yes. The fact that you're not doing anything explicit and you don't have an an external record of the message digests isn't actually being asked in question 30.

D [02:34:11] True. So I suppose I guess maybe that's where we come back to the range again, then some people might say that was enough and would happily say, yes, our collections are being, that's happening on our collection, other might take a slightly different view, I suppose. So that's whether the range of between the five percent and the 95 percent starts to come in again, I guess, on this one.

E13 [02:34:37] Might it be helpful if Question 30 were clarified to to focus on an external record of file fixity as opposed to relying upon the storage provided, for example?

H [02:35:11] I don't really have an opinion, Alex or David?

D [02:35:18] I think that's probably really what we were after, I mean, we've got the thing about integrity there and the assurance, the assurance is identical. So you actually having, demonstrable evidence of that, I mean, that that could in part be logs from from ZFS or something showing the ? happened or whatever. But if you'll just purely trusting it to get on with it and not doing any checking is probably not sufficient.

E13 [02:35:45] The the activity that I personally would be interested in is the maintenance of an independent record of the file fixity.

D [02:35:57] Yes, also also important, yes. During.

E13 [02:36:01] And that that, I think, should be the the purpose behind question 30. It just needs the wording of the question.

D [02:36:15] Yes, because otherwise, yeah. The ZFS type type set out that the checksum is there, But isn't isn't externalised and the same with with cloud storage as well. So yeah. Yeah. I think that's, that's what we were driving at. Yes. We should have made that clearer.

E13 [02:36:31] Thank you.

H [02:36:39] Alex, can I ask your opinion. So we're getting to one o'clock now. We had planned to have lunch between one and two. Do we think it's best that we pause and resume continuing questions after lunch or carry on for a bit longer? And I know (name)'s got to go and feed children.

A [02:37:01] I think we should have a break. Yes, let's have a break. I think would be useful because we have got 44 questions. I don't think we can carry on at this point. I think it would all do us good to have a break. Yeah, I think we should keep to the one o'clock. Thank you, (name). And come back at two. Do you think?

H [02:37:23] I think that I think that's good. So we'll will pause, we'll resume these questions at two o'clock. So hopefully you all all free at 2 to continue. And we'll continue from two until we finish going through the questions. And it'll be at that point, we'll then leave you to go and answer them in your own time.

H [02:37:42] Yes. If you can't be rejoining at 2:00 or or your difficulties, we all we will be recording that and we'll record a separate file. Or when you're answering the questions, if you have any questions for clarity, you can message us. Or e-mail or call us and we can explain anything to you. So we'll pause now, thank you. We are going to end the meeting so that we have a time to save this recording ~~and that if I stop the recording for now.~~

[02:38:12] ~~So.~~