Keynote speaker: Digital salvage operations — what's worth saving?

Hjalmar Gislason, vice president of data, Qlik | Oct. 13, 2016 | Charles E. Young Research Library, UCLA | Dodging the Memory Hole 2016: Saving online news

---

HJALMAR GISLASON: ***[00:07]*** So, as Martin said, my name is Hjalmar Gislason, and that's probably the most complicated part of the presentation. That name comes from Iceland, so that's where I grew up and that's where my funny accent comes from, and actually some of the stories I'm going to tell you are from Iceland as well, but they definitely are relevant in the bigger context.

[00:28] So with that, what I'm going to talk about is kind of what's worth saving, and I'm not necessarily going to answer that question, but…

[00:48] Like I said, I'm not going to try to answer that question, but maybe kind of get some of you guys thinking about what is worth saving and what is not worth saving, and do we want to save everything. And I'm going to start it off by a little story. So, this is Deaf Teddy. This is actually not a picture of Deaf Teddy himself, but Deaf Teddy was a teddy bear that my wife had when she was a little child. This was her favorite toy. This is the toy that she slept with every single night. This is the toy that went with her everywhere. We all know kids that have toys like that.

[01:24] Now, Deaf Teddy was called Deaf Teddy because his ears were missing, and he was worn. He was old. He had been owned by somebody else before my wife had him, so he was not much for the grownups to look at. So, at one point, my father-in-law is doing one of his few household chores, which is kind of tidying up in her bedroom. So, he sees this old teddy bear lying around, and he throws him in the trash and he's gone forever. There was, as you can imagine, a big and sad event in the household, and still to this day not forgotten.

[02:10] Another story of my father-in-law is that — I come from a country called Iceland, but the name is deceiving. Sometimes the winters are a little bit milder than we expect. One winter — this is after I met my wife, this is probably about 15 years ago — there was very little snow, so we didn't have a chance to use our ski equipment that we were storing with my parents-in-law. The next winter, when we come there and we're going to kind of bring out the ski equipment, we can't find our ski boots and we ask my father-in-law, and he said, "Oh, I threw it away. It was never used," because it hadn't been used for more than a year now because there was no snow. So, some people are like that.

[02:56] They want to throw everything away and they don't like a lot of things in their life. My mother-in-law, on the other hand, is the exact opposite. Old knitting projects from when my wife was in school are still lying around. She knows where they are. She knows when they were worked on. She knows everything. She kind of tries to hide things away from my father-in-law so that they don't get thrown away and that's kind of the balance of the household. That's probably kind of how it all

works out.

[03:22] So, the question that we really want to answer is: do you want to be more like my mother-in-law, saving everything, storing everything, not throwing anything away; or like my father-in-law that only has these few, absolutely necessary things in his life and would like to get rid of everything else? Now, just a little bit of a background on me. Martin kind of covered this, but, in short, I have this title VP of data at a company called Qlik. Qlik is a visual analytics company, fairly large in that we're in 50 countries with about $700 million in revenue.

[04:02] My role there is on the product management side of the house, where the title VP of data basically refers to that I am responsible for everything that happens before our customers start to visualize and analyze the data. So, how do you read the data in, how do you prepare the data before you can visualize it, our big data effort, data management efforts and so on. This is all on my plate. I joined Qlik two years ago through the acquisition of a company I started called DataMarket. It was started in Iceland, but in 2012 I moved to Boston where I currently live, building out the business part of that.

[04:49] And before that, I had another startup called Spiral.net, which was actually about allowing people to save what pages that they were visiting, save copies of what pages they were visiting so that they could refer to the exact copies that they saw. So, I've kind of been on the data side of the house for quite a while. My other interests, as Martin mentioned, is that I'm in media. So, I've never worked in media, but I'm a news junkie. After I sold my company I had some discretionary budget, so I invested in this company called Kjarninn, or "the core," which is the literal translation of that, which is the closest thing we have to an investigative journalism institution in Iceland. And then Martin mentioned the New England Center for Investigative Reporting.

[05:34] So this here is the future-proof chart of data. We used to have a lot of data. Now, we have gigantic amounts of data, and soon we will have gargantuan amounts of data. I've seen this chart often with some numbers on them, but it changes from year to year and nobody agrees, but this is future-proof. This will work forever. The point of it being that there is evermore data, and we've all heard this story, and its data of many different types and there's data coming out of more and more human endeavors than ever before.

[06:13] I think that one of my purposes today is to kind of broaden the perspective a little bit on what we have to think about when it comes to preserving data. So, you have the web. We all know the web. We have the deep web that I will talk a little bit more about, and then you have all the other data out there. And, where you might look at news content as something that definitely setting up in the web part of this, and just a small part even of that. I hope I can convince you that there are things that you need to think about, at least in the context of news, that are definitely not a part of that top of the pyramid.

[07:00] But we start with the web itself. This is your turf. I'm not going to be the egg trying to teach the chicken or the hen here. So, you know all these different media organizations, how they work, you know how they publish, the publishing schedules, how they've been changed, how the delivery mechanisms have been changing, how the consumption models are ever changing and so on. Earlier this year — so this is our former prime minister of Iceland.

[07:30] He was a pretty prominent figure in the Panama Papers that were leaked earlier this year and the International Consortium for Investigative Journalism did a really good job at getting a large group of journalists all over the world working and digging out some of the interesting things in

there. The company that I'm involved with, the media company, "the core," Kjarninn, was peripheral to that work. We were not the Icelandic partners for that, but I got to kind of see a little bit into the process. But what I want to talk about here is actually more about the consequences of that, and what that means for media. So, on Sunday, which is probably the third of April, the Panama Papers come out and all of the big stories are published simultaneously all over the world.

[08:33] The chain of events that followed in Iceland was so intense that the amount of things that were going on, the amount of meetings that were happening, the amount of news content, newsworthy things that were happening were almost more than you could have covered in real time. So, you know there were some journalists covering something in one place, other journalists covering something else in another place, and especially the Tuesday, which is I think the fifth, there were...well, first of all if you think U.S. politics are a soap opera, this was definitely one, too.

[09:14] The point here is that certain media publications were covering parts of the story, and what unfolded and how events played out will, I think, be something that people would be really interested in looking back at. So, my point here is, it's not enough just to have saved all the stories that came out, but you almost need to know exactly at what point the story looked in a certain way, who published what before somebody else and kind of chained that altogether, because as things were kind of unfolding, as you want to look back at this day — even just the time stamps of the news aren't necessarily something to be trusted, and there are all sorts of things. So, basically, on a big news day in order to properly capture the news, it's not just enough to capture every single story. You almost have to capture all of the media on a minute by minute basis so that you can play back the way things went.

[10:19] So, that's kind of a takeaway from this story. It's not an Icelandic phenomenon. So, you see other things changing as well. It goes back to, you can't just trust that you can go back to a story at a certain point in time and trust that it hasn't been changed. So, the name of this event, Dodging the Memory Hole, wasn't lost on me. The memory hole in George Orwell's "1984" is where news went to be remembered forever, meaning forgotten. The job of the main character in the book was actually to take news, historic news, and change them to the liking of the rulers of the time.

[11:09] So, this is a pretty famous thing that happened at The New York Times earlier this year. It was a relatively pro-Hillary piece that came out that was changed later on to kind of be, well, changed in tone of voice, and somebody pointed out this difference in the news. It actually turns out — and this comes from a service called NewsDiffs, if you haven't heard of that, they do exactly this. They track the same piece of news over time and check out the differences, and it turned out that this particular article had changed quite a lot.

[11:53] So, once again, you can't just capture an article as it stands and trust that it will be like that forever. You almost need to capture the same article over and over again to make sure that it hasn't been changed. Martin and I were actually talking about it earlier today, that even the news that are there, even the articles that are there — so, let's say that you reference CNN.com 10 years ago. The piece of news is probably still surviving on the same URL, but it's obviously now probably in a different CMS system.

[12:28] It's in a completely different context on the webpage, and even if the text of the story hasn't changed, some of those things can actually be important, too. So, there are also other things to think about and consider here. There are obviously a lot more things to be said about the web, but I want to move on to the next layer, the deep web. This is a thought that came from, about, I think it was about eight years ago. I was lucky enough to be invited by the Icelandic National Library to talk about their future strategy about eight years ago.

[13:12] This is when social media was very new. Facebook was very new. But, still, people knew what it was, and they started to understand that it was an important medium. The picture there in the middle is actually of Iceland's only Nobel laureate. So, this is a guy called Halldór Laxness. He won the Nobel Prize in Literature in 1953. My question that I posed to them was, "Had Halldór Laxness been on social media as a teenager, would we like to have his social media content as historians or librarians?" The answer is probably, yes. That would have been interesting. Would he have liked that? I don't know. Would it paint him in a different light? Would it have been ethical to capture all of those things? And with social media there's so much more things that are coming out about people now.

[14:19] But, you know, the people that are growing up, even — let's just take the, what was it? Was it a 10-year-old or 13-year-old — no, 11 years since the tape that has been chasing Donald Trump for the last few days came out, and that's in an era where you actually had to have a camera crew following you in order for something like that to be captured. Think about somebody in that situation today with all the kind of capturing on social media, all the video, and all the imaging, all the recording that's happening all of the time. So, politicians, whoever, they will have a lot more things in their recorded history in the future than they have now. The question is, how easy is it to get to that? What's ethical to say? What's ethical to bring up? What is private? What is public? And so on. These are all kind of interesting questions to think about.

[15:24] Another aspect of the deeper web is the fact that we as archivers can't really get to the content. Some of it is hidden behind usernames and passwords and you have to log in and you have to have access to those to be able to kind of capture them. But others are just trapped in proprietary systems that our crawlers or our indexing engines don't know how to read or how to get into or how to query for all the content that is there. They are also often really large databases that we are then trying to replicate in a less ideal format somewhere else, and there are also the problems with that.

[16:06] And, you know, Facebook is a very good example of this, where a lot of the content, actually most of the content that's published there, is not public, meaning it's relatively public. It's published to all of your friends or some groups of people, but it's not something that — first of all, Facebook makes it hard to archive it. Secondly, it's really big. And thirdly, a lot of it isn't for everybody's consumption. So, Facebook is really the only company, or the only organization, that is in a position to even consider this. And I mean, they obviously archive it.

[16:44] They store it and we can reference back to it, but we have no claim to be able to see that. Especially, because a lot of news these days, as was pointed out earlier, comes from these types of sources. When you're thinking about preserving the news, then, for example, social media profiles of at least public personas, people that are in the public sphere, are they a part of preserving the news? Is that a part of preserving the news as well? Because, even something they say now might not fly, but it might become really newsworthy at some point later on.

[17:29] So, again, there's a lot more to be said about the deep web, but I really want to spend the rest of the time here talking about all the data that's out there. That's probably mostly my, at least in my last two companies, DataMarket and now with Qlik, that's what I do most of the day. So, YouTube. Anybody care to guess how long it takes for 500 hours of video to be uploaded to YouTube? A month, a week, a day, hour, minute, second? Well, it is a minute, so you're off by a factor of 60. But that's still a lot of content.

[18:17] It means that it would take three man-months, three person-months, to watch the content that's uploaded to YouTube every single minute of every single day all year round. So, obviously,

nobody will watch all of that. Meaning no organization will go around watching all of that. Probably most of the content is watched by somebody at some point, but not even all of it. So, what do we want to do with that? There's a lot of newsworthy content out there.

[18:57] But we have no clue which parts are newsworthy. We can go off public profiles. It would capture some of it, but what about somebody capturing an event that's going on, published on YouTube, [it] doesn't fly on social media, surfaces a year later, or maybe never surfaces, but could bring a new perspective on something that's worth knowing. And I definitely don't have the answer to that. But it is something to consider and something to think about.

[19:28] And on that note, I want to tell you a little bit of a story about data gathering. So, this is a place in the western part of Iceland, called Hornbjarg. It's a lighthouse in the middle of, as much in the middle of nowhere as you can get. And so, this is Iceland. Iceland is the middle of the Atlantic Ocean, and up there we have Hornbjarg. Almost everybody lives here in the Reykjavik area, and then kind of around the coastline, and up there, there's only this lighthouse. Now, these are the people that were the keepers at that lighthouse for about 65 years.

[20:10] And one of their tasks, until the light keepers were replaced with automated equipment, one of the tasks was actually to measure the weather. And one of the measurements they would take would be the temperature. So, on May 12, 1960 at 12 o'clock, 4 degrees Celsius. Then he goes out again six hours later and he takes another measurement. Temperature is obviously just one of the many measurements he's doing. This is May 1960, every six hours, actually later every three hours, he goes out and takes all these measurements. This is over two years. You can see that it was slightly warmer in 1960 than what it was in '59. This is a whole decade and this is just one of 200 such manually-manned stations around the small island of Iceland.

[21:11] Now, imagine all the man years that have gone into gathering this data that can be really interesting for scientific reasons, for historic reasons, for even news reasons; to go back to and study what the weather was like. And this is just the weather. There are all sorts of other people that have, either as a big part of their job or as a kind of outcome of their day-to-day job, some sort of data gathering. And a lot of this data is lost. Fortunately, most of the weather data, actually all of the weather data that was captured in Iceland, has been found. So, we found the outcome of these man years of work, but it was by coincidence. There was nobody who was really thinking about archiving this as it was coming in. So, just something to think about, preserving data is actually about preserving the work of the entire working life of people all around the world. And it would be a shame to lose that, especially if it can be helpful to do research later on.

[22:35] Another piece of data that, when I started DataMarket we started digging around for, was earthquakes. So, earthquakes and eruptions are fairly frequent. Iceland is a volcanic island. And we have quite a lot of measurements around that. And there turned out to be an archive of all the earthquakes. It was not easily gotten out. But we started looking at what could be done with that data. And this is maybe more to point out that it's not — well, it obviously helps archiving the data — but then the ways you have to look at that data changed quite a lot as well. You may remember that a few years ago everybody that was traveling in the Western Hemisphere more or less were stuck for a week because of a volcanic eruption in Iceland. And, this shows the chain of events there. It so happened that we were working with this data at the time and created this video.

[23:52] So, the way the video works is that you have camera circles that represent measured earthquakes. There are, on average, about 250 of those in Iceland per day. So, there's quite a lot of them. Most of them far too small to be felt, but they can be measured. And the size of the circle shows the scale or magnitude of the earthquake, and the color shows how deep they are. So, if we

play this out, we can see earthquake activity here in the early 2010 glacier, so speaking of hard to pronounce names, this is [Icelandic]. And you can see kind of scattered activity all around the glacier for a while and it starts concentrating.

[24:40] And you can actually see how some of the activity moves around, concentrates, and then you can see as it starts moving to the east. And this is actually, you're actually literally looking at the magma finding its way through the ground and you will notice that they get ever shallower as well. And it ended up that the magma found its way up through the ground between the two glaciers there. And the nature of volcanic eruptions is that if they happen on land, they just become lava and they are relatively contained. If they happen under glacier the magma turns into ash in a big explosion immediately, and that's when you get these big ash clouds. So, this was a small and beautiful eruption. A lot of people went up there on their Jeeps just to look at it, and it went on for about two weeks, but the activity didn't stop.

[25:36] So here we have a certain activity and the big eruption that happened at the top of the volcano. So, the point here is to — this video actually ended up being licensed by the National Geographic and they created a documentary where this was featured — but it's just to show that if you don't have the tools to look at the data in context or in an informative way, you don't even know what the data is telling you. So, preservation is also about making the data available, and then it is our, or your, task to also provide people with some of those tools, or is that their tasks entirely; which obviously means that there's a very small part of the population that knows these tools or has the ability to kind of get to them in that way. Another example of such data is it turned out...

[26:48] So when the crisis in Greece happened, the economic crisis in Greece started unfolding, one of the things that surfaced was that the national statistics that their acceptance into the European Union was based on turned out to be wrong. Some say falsified but it was at least not correct. So, there were certain measurements that you have or standards that you have to meet to be accepted into the European Union in terms of inflation, in terms of unemployment and things like that. And it turned out that those were off and they, at least they wouldn't have met those if the statistics were correct.

[27:27] The interesting thing here is that the data that was published at the time, which is published electronically on the website of Statistics Greece, was nowhere to be found, meaning there was now a new database out there with the correct data. That's what you would get if you go to the website and look it up, but there was no digital record of the data, the way the data looked like when it was published five years earlier. And that goes to show that, first of all, the website was probably archived by the Internet Archive and probably some others as well.

[28:10] But they didn't see into the system. They didn't see into the data that was being published on the website. And the only way to recreate that was actually to go back to things that had been printed and recreate the data series from there, which is interesting and obviously not something you would do except in extraordinary cases. So, a lot of this is about, how do you trust that the data you can get now and the data that is at a URL is the same as it was before? This is very much what we were doing at DataMarket back in the days. We were gathering all sorts, mostly time-series data. Our slogan was, "find and understand data." So, we helped people find. We had a Google-like search that searched through databases of statistical data and then surfaced that in a uniform way.

[29:04] So, you could go in, type in search queries, and then we would find things like population connected to urban wastewater treatment. United Nations electricity production, this is from the U.S. EIA. Transport of passengers, this comes from the Europeans Statistics Office and so on. So, we were gathering this type of data from about 250 organizations around the world. We had 150,000 data

sets, and we were capturing the data in data format, in tabular-structured format. It covered about 6,500 years of historical data, about a billion time series in there in total, and about 8 billion facts. So, a fact would be something like the temperature at Hornbjargs at twelve o'clock on May 1, and so on.

[30:00] So, quite a lot of data there, but it was only a very, very small fraction of all of this type of data that exists. So, one of the reasons I got kind of involved in these types of things was actually that we were working on a research project with a few research organizations on how to preserve exactly this kind of data. Kind of how to go one level deeper into the types of systems that publish this, because we had about 250 organizations that we were capturing data from, and we probably had 249 different ways of accessing the data. There were Excel sheets, they were differently formatted; there were proprietary systems, even the proprietary systems that were using the same systems, they were differently configured and so on. So, there's a lot of things to think about there. And, again, this all relates back, all of this — there's a story in every single piece of data, and to kind of prove that out I have a couple of my favorite stories here.

[31:01] So this is electricity use in Iceland, in gigawatt hours. The blue line is household consumption, and the orange line is heavy industries. So, with mountains and lots of water and lots of geothermal energy we are a somewhat resource-oriented economy. You can see that back in the late '90s, we had the heavy industries at a certain level, and you had the household industry at a certain level. Even just the big picture here shows you the growth of heavy industry in the two decades that the data covers here. But there are other, more subtle factors here and it goes back to — you have to think about all sorts of things when you're looking at the data. So, you notice these regular dips here.

[32:04] So, these are mostly aluminum smelters and big factories that go 24/7. You don't want them to fluctuate and so on. So, there is obviously a regular dip here. Once a year, the energy consumption drops significantly.

[32:21] Anybody care to guess why that is? Christmas? Summer? For the household, this is the annual swing. So that's what you're seeing there, you're seeing the seasons. If I tell you this is February, does that give you any hint?

[32:48] No, February has 10 percent fewer days than the other ones. It has 28 days on average, 29 in some cases. So, it just goes to show that you have to be very careful when you are looking at data. I always laugh when I hear somebody point out that some number was, let's say credit card spending, was point 5 percent less this year than the same month last year. OK. How many weekends were there in that month this year versus last year?

[33:25] So these types of comparisons have to be really carefully considered. Now this one is not.

[33:37] This one has some strange characters in it. So, this is another data set. This is to show you that by bringing together different data sets you can tell a story as well. So, these are, translating what it says up there, these are actually Icelandic Talcos and their prevalence in their market share. This is an age pyramid. And the gray is showing the overall age pyramid. The overall age composition of the country and the colored parts are showing their customers. So, you can see this is the old incumbent; mostly old people or older people.

[34:17] This was a Vodafone goes for like a family strategy so they have people in all ages. Tal is another one, similar strategy in the same shape but just smaller. And then this was the, kind of,

newcomer that went after the youth market. These are people that do not have cell phones. So, this is male. This is female. It's only old women that don't have telephones. And then the last one is showing something else. So, the story here is by combining data about the actual research that was done by Capacent, which is a Gallup representative in Iceland, and the statistic from Statistic Iceland, you get in a single picture a very easy to understand overview of that market. The last story I want to tell you, so again, yet another Icelandic story, but this is the population of Iceland from the early 1700s.

[35:24] So we've had our share of diseases and eruptions and people moving abroad and so on. You can see some small dips here. There is about 15 percent of the population that died off in the late 1700s. So, these are significant but you can also see that for some reason I'm leaving room up there. This is the number of sheep in Iceland. You can see that they outnumber us by a factor of five on average. But there are still some interesting things here. So first of all, the data isn't quite as frequent to begin with, but there's also this one point in time where we outnumber them, and that was in the beginning of the events that were happening here in the late 1700s where, well, it was a big volcanic eruption once more and it killed off more of the sheep than it did of the people. So, for a short while we had the upper hand on that.

[36:41] Actually, that brings me to one more point. So, I was thinking this morning — if some of you are on Twitter and you've been watching the hashtag for the event, you might have seen [that] I tweeted with the hashtag and I was wondering, is it more valuable, do we learn more from reading today's newspaper, I'm going to say a newspaper here, or a newspaper from 30, 50, 100 years ago? There is no definite answer to that. But the reason I started thinking about that, and for some reason all of my stories are about volcanoes, is that somebody dug up a 40-year-old newspaper article.

[37:30] It was actually a letter to the editor of the biggest newspaper in Iceland. There had been a big eruption on a small island south of Iceland. It had a population of about 5000 people and they had to be evacuated overnight. So, they were all coming to Reykjavik where there was a housing crisis because the economy was going quite well and they didn't keep up building enough houses there. So, the letter to the editor literally says, "we can't have all these people from Vestmannaeyjar," which is the small island less than 100 kilometers away from the capital, "we can't have all these people come here. There's no room. We have to build for our own people first." And I thought that was interesting because that's exactly the same conversation that's going on today.

[38:21] But the people are just further away. They are in Syria. They are in northern Africa and so on. What it brought home to me was, you would never say that today about people that lived 100 kilometers away, but we're happy to say it about people that come from ten thousand kilometers away. So, it was just an interesting lesson in reading an old newspaper, but also in the context because just browsing through the newspaper you see what somebody found worthy, pointing out, writing up and printing on a piece of paper at that point in time. So, no matter how seasoned you are in history you will learn something new from reading that. Now, in the born-digital age, what will that experience be forty years from now? How can you go back to capture the spirit of the day, if you want, forty years from now?

[39:21] Which is — and I'm not saying one is better than the other — it's just, how would we recreate that? Even with full preservation of all digital artifacts, how do you get to that experience, or is it possible? Anyway, the final piece here is about, there's also data coming out of more and more things. The Internet of Things is a big thing. It's a big kind of hype word but people say that there will be — the machine data connections will outnumber the human data connections by a factor of at least 10 if not more than that. Light bulbs are gathering data about the energy consumption when they're on and so on; cars; all sorts of, basically all of our things are already gathering data and a lot of that data will be coming online. And is any of that worth saving?

[40:20] The interesting thing there is that you have this kind of flow of data, and there's no way, even with the storage capacity that we have today, and the cheap computing power and so on, there's absolutely no way to store all of this data. It's impossible. But we know that some of it we want to store, some of it we want to analyze, and so on. So, it's going to be super important in that world to have equipment that can take a split-second decision about what to store and what to throw away. Most of it we're going to discard. Some of it you want to store, and some of it you want to analyze in real time; [if] something interesting is happening and so on.

[41:10] So this is a problem in industry, but it will probably be a problem — and you could throw this to different types of skills. This is in general my father-in-law or my mother-in-law question, what do you want to do with all the data that's coming your way? What is important? The surprising thing is that you never know what will come out of these types of things. So, one of the more interesting things that are happening in the computer world or in the software world these days is artificial intelligence. This is from earlier this year when a computer beat the best Go player in the world. Chess had been conquered before and so on. So, artificial intelligence is becoming super good at certain tasks.

[41:58] You're beating humans in ever more concentrated small tasks, like playing chess or Go, driving a car, reading an X-ray photo. Those types of tasks computers are becoming or, if they haven't already become, better than human experts are doing. The only reason for that is that it can build on a lot more experience than a single human being, meaning it has access to all this data, [and] trains it [to] become good at these tasks. So, a lot of land grab is going on now in who has the best databases of these types of sorts. So, in order to build the best self-driving car, whoever has the most data about driving cars will win. And about the kind of geography and about the landscape. You will see unexpected uses for that.

[42:55] If we go back to, let's take an example that I know quite well, let's say that you take all the fishing that happens around the island of Iceland. If you had all the data about every single trawler and where they were, what's the temperature of the ocean, how long did they go for, how much fish did they capture, what were the fish they captured and so on. If you could bring all of that together, you could train a neural network to be better than an individual captain at going after the fish and fishing for it. Those types of things will happen, and people use good archives to do these things, and some of these things will be super valuable. You just don't know which things yet. So, the takeaway here is there's more to archiving and news archiving or otherwise than just the web itself. We didn't even talk about mobile apps and other things where the digital content is very hard to get at. Hoarding is not a strategy. You can't capture everything.

[44:03] You have to have some rule of thumb in what you keep and what you throw away. And thirdly, you can only guess what will be important in the future. It can be very surprising what will be important in the future. So, this is from a newspaper in 1927. This is from the front page of The New York Times. I don't remember the exact date, but in 1927. It's the third sub headline on the same main headline in the paper; and anybody care to guess what this is about?

[44:43] You got it: television. So, they had the president appear in the first broadcast television. He spoke to people across that. You can just imagine how amazing that must have been. Bet they were like, "Hm, how is anybody going to use that?" I mean it's cool, but does it have any application? So, it's just to point out the fact that we can only guess what will be important in the future. So, these are the three things that I want you to take away, and I've blown away most of my time if not more than it. Do we want to take some questions? Three questions. OK.

[45:44] OK. Yeah. So, repeating the question, so essentially the kind of hoarding is not a strategy resonates well with the librarians when they're faced with this, you know, even with paper. And now you have more and more data coming your way. Is there a way that — it's an overlap between data science and librarians. What could that relationship look like?

[46:09] So, first of all I think you're absolutely right. The best, but far from perfect solution, is probably to try to bring data science into that somehow. I would dare to say that probably just because of funding issues and other things like that, industry may be a little bit ahead in some of these things, so I would look to a lot of the strategies that have already been applied in industry and see to what degree they apply there. The problem with these things is number of skilled people and how costly they are.

[46:50] So that's really the big problem because I think that it's like any kind of cross-disciplinary mind meld; if you have people that know what they're doing in two different areas or functional areas, something great can come out of that. I think it's more about — the problem to solve is, how do you get to the data scientists and how do you get them to listen and take notice of your problem? Because as soon as they do, I think that even in just the very first things that could come out of that could be super valuable and they could probably be done quickly and relatively cheaply.

[47:29] So it's about getting the attention of the right people and getting them together in a room for a day. That's not an answer but it may be a method.

[47:41] It's a super interesting question. I'm not going to claim I have the answer but... Yes, so summarizing the question, he's going to be talking tomorrow about the Vietnam War and the archives of some of the news coverage there. But there have also been tens of thousands of books written about this subject. Probably taking different parts of it or summarizing and bringing some of those things together.

[48:06] But how could you potentially integrate some of the overview material with the news coverage of the day. Is that a fair summarization? Yep. So, one of the interesting things is that obviously perspective always comes with time. So, you need a certain amount of time to start to console it and start to really understand what's going on. And you get additional pieces of information there, which is often backed up with this. With archives and with the possibility to preserve a lot of this, I believe or I imagine that some of what you will be talking about is that we were at risk of losing some of the base material or at least it wasn't easily accessible.

[49:00] So it's the same problem as Martin and I were talking this morning about [with] scientific papers. So, scientific papers are archived in a fairly good way, but everything they reference, especially if it's online, is a problem and all the data they reference and all the data they publish as kind of supplements to that is a problem. So, I think that it's a little bit akin to the same thing. And you know now that everything is digitally born. So, obviously first you have to solve the problem of how do you archive it at all. But you will definitely see new possibilities in the fact that you can, I don't know, actually I think a book will look very similar in 40 years to what the book looks like today but you will have the possibility to reference all sorts of supplementary material and assume that people can get to it.

[50:52] So, the difference there being that now, or at least 20 years ago, when you published a book you would be referencing things that you would have to dig hard in a really good research library to be able to even get at what's being referenced. But, you know 40 years from now, hopefully with a good archive, you can assume that if somebody wants to take the deeper dive on something you're referencing in a piece, an overview piece, that they can get at it, and the experience will be different.

I imagine that we may jump more back and forth. I find this with myself when I'm reading a book, I find myself picking up my phone and googling for something because I need more depth or want more depth from what I'm reading about.

[50:36] I'm sure that will be almost built into any such overview literature that we'll see in the future. Last question. So, the question is, are there any centers of concentration or kind of excellence on some of the topics that are being discussed here. And, I think we're it. I mean partly. So, the IIPC conference that I was at in April is definitely probably one of the higher concentrations you find of people that are, A) thinking about the problem and B) have some of the skills to do that. I think what is needed is that we need more people to come from the technical side of things to just pay attention to this problem, and start applying some of the things they know there.

[51:29] And the good thing about us nerds is that we are often easily pursued by other things than money. We take interest in weirder things. I think that, so for example, I don't know if you know of a group called Hacks/Hackers. It's a meetup group that happens all over the US and probably in other places in the world as well, where the number one thing it's about is it takes the hacks, the journalist hacks, and the hackers, the kind of the computer nerds, and just brings them together for beer. You know, that's more or less the purpose, once a month. But there's always a speaker. A lot of really interesting data journalism stuff has come out of that.

[52:16] So I think this type of outreach over to the technical side is probably a little bit akin to the first question. It's like, how do you get those people interested in the problem? Because as soon as they do things can happen with a modest amount of money and sometimes not so modest amount of work but they're willing to put it in. So, we're way over time.