Panel: The future of the past: modernizing The New York Times archive

Evan Sandhaus, Jane Cotler and Sophia van Valkenburg, The New York Times Technology Team
| Oct. 14, 2016 | Charles E. Young Research Library, UCLA |
Dodging the Memory Hole 2016: Saving online news

EVAN SANDHAUS: ***[00:08]*** Thank you for bearing with me while we get the technical things ironed out. I'm Evan Sandhaus, and I am here with my team today to talk about some work we've been doing with migrating The New York Times archives. A little bit later, once we say our piece, Ed is going to interrogate us and ask us some questions as a panel.

EDWARD MCCAIN: And we'll open it up for other people, too.

SANDHAUS: Absolutely. Just to get things started, opening formalities, the three of us work for a news organization called The New York Times. It is arguably among the most significant English language news organizations on the planet. And if you haven't heard of it, I totally suggest you check it out. If it's a valuable part of your daily ritual, why don't you consider becoming a subscriber? So, today you've got not one, not two, but three folks from The New York Times. Myself, I'm an executive director of technology, with oversight of search, archives, taxonomy and a few other things, but those are the ones that are most interesting for this crowd. I also have with me, Jane Cotler, who can introduce herself, and then Sophia.

JANE COTLER: ***[01:28]*** Hi, I'm Jane, I'm a software engineer on our discovery services team. In addition to the work on the digital archives that we're going to be mainly focusing on, I work on two products called TimesMachine and Switchboard, both of which you will also hear about today.

SOPHIA VAN VALKENBURG: ***[01:49]*** I'm Sophia, I am also a software developer. In addition to our archive work, I also maintain a back-end web API for The New York Times.

SANDHAUS: ***[02:01]*** So, we're going to talk today, and then have a discussion about, a problem of archives; a problem that, I think, almost everybody in the room is familiar with. First of all, I can't say how exciting it is for me to be in a gathering full of people with whom I can discuss these issues with no preamble and lesser fear of boring my audience. This has been a real pleasure. A second thought I had about this gathering — this is the Dodging the Memory Hole gathering.

SANDHAUS: ***[02:29]*** And Orwell also said, "He who controls the past, controls the future." So, look around. You control the future. So, the problem of archives, that we have to deal with today, that we are dealing with at The Times and I suspect a lot of you are dealing with, which is: How do you faithfully represent information created with one technology, using another? This is, I would suspect, a familiar problem. It's something that we've dealt with in many, many different ways throughout our

history at The New York Times. And I want to talk about two examples, and then hand it over to Jane and Sophia to talk a little bit more about more recent work in this. One of the first examples I could find of The Times needing to migrate information developed with one technology to another one…

SANDHAUS: ***[03:22]*** Actually, you know what, I'm getting ahead of myself. Before I start talking about this, I want to just, sort of, level set about what we mean when we talk about The New York Times archive. The New York Times published its very first edition on September 18, 1851. And since then, we've published approximately 60,000 complete issues spread across about 3.5 million pages, and those pages contain about 15 million articles. Now, that is the print archive. Our digital archive is divided up into several different tranches, as well. The digital archive is primarily divided into two main periods.

SANDHAUS: ***[04:17]*** Basically, before 1980 and after 1980. What happened in 1980? Well, in 1980, The New York Times decided that it was probably a good idea to start persisting the digital text of the news that they were creating. I believe there was a digital production process in place before that time, but we did not persist that data. So, from 1981 onwards, we have the full digital text of our archives. Before that, we have less stuff. Like I said, one of the challenges we deal with, is, how do we take stuff that was developed for one technology, and present it using another one? One of the first examples of that was a system that we announced in 1968, and then launched to production in 1972.

SANDHAUS: ***[05:10]*** And that system was called The New York Times Information Bank. This is actually an old problem for The New York Times. The data in question, in this instance, was The New York Times Index, which we had published since 1913 as a bound, digital index of the contents of The New York Times. It is, and remains, a very valuable research tool, but the fact that it wasn't digital made it pretty hard to search. So, we decided that we would digitize that, and in that time period, we did. I was supposed to be skipping these slides, I apologize.

SANDHAUS: ***[05:48]*** That system, when it launched, provided the ability to search our index digitally for both the newsroom and for outside clients. This is a picture of folks in the newsroom using this technology in the 1970s. So, that's one example of how we've had to take old technology and make it new again with new technology. A more recent example is an effort we've done called TimesMachine, which is, essentially, a virtual microfilm reader. This is one of my favorite headlines The New York Times has ever published. It was published 104 years ago today.

SANDHAUS: ***[06:25]*** And it is: "MANIAC IN MILWAUKEE SHOOTS COL. ROOSEVELT; HE IGNORES WOUND, SPEAKS AN HOUR, GOES TO HOSPITAL." It's a spectacular headline. I mean, my condolences to the president; and, of course, it's bad news, but still, I mean, wow. So, the TimesMachine was sort of conceived to address a challenge with that deep archive. As I mentioned, everything before '81 we didn't really have text for.

SANDHAUS: ***[06:53]*** So, it wasn't born digital, and we weren't in a position to serve it digitally the way we would more contemporary stuff. And that was an extra challenge when you look at the vast sweep of our archive — that little graph is the number of articles we've published by year since 1851 — and you can see that about 75 percent of the creative output of the newspaper was published before we had digital text. We call that part the deep archive, and it consists of about 46,000 issues, 2.3 million pages containing about 11.5 million articles. And that scanned archive consists of some pretty cool material. We have scans of all of the microfilm. Gosh, I wish it was scans of the original newspaper, but it's scans of the microfilm.

SANDHAUS: ***[07:39]*** We had that microfilm segmented, so we know where on the physical page stories occur and we also have some metadata. We have the headline for every article. We have the lead paragraph for every article. And we have, what should be familiar to you, and what we affectionately call, the Dirty ASCII, which is the nasty OCR text that is good enough to build a search index, but you wouldn't want to show somebody. And that's an example of why, right there. Also, we have that indexing metadata. We have subject headings and abstracts for almost every article in that time period. So, given that, we were like, we think this is the right set of ingredients to build a really cool experience for our users around our archive.

SANDHAUS: ***[08:19]*** So, I want to sort of make the point about how we've changed things in our archive. I want to discuss with you John Fairfax for a second. I don't know if anybody in here is familiar with John Fairfax, but he was the subject of one of The Times', I think, most spectacular obituaries ever. He passed away in 2012 and his life included an apprenticeship to a pirate; he once almost died at the hands of a jaguar; and his claim to fame was that he was the first person to row across the Atlantic solo.

SANDHAUS: ***[08:57]*** So, if you wanted, prior to the advent of TimesMachine — and I should mention there was an earlier product called TimesMachine, as well. But, prior to the advent of the most recent edition of TimesMachine, the way you would have found this article in our archive is you would have done a search, and you would have gotten a link to that article, and it would have taken you to this abstract page; which gives you the headline, that lead paragraph that I said we had, and then contains a call out, this stuff didn't used to exist, but you did have the opportunity of seeing the high-resolution PDF of that article. So, what you get of that article is a scan of Mr. Fairfax, waving from the vessel in which he rode across the Atlantic, and the story, the break and the rest of the story.

SANDHAUS: ***[09:41]*** And that's cool. It's really cool to be able to see that original piece of coverage. But, here's the thing: here's that same story in TimesMachine. Let me zoom in on it. Same guy. Same wave. Same story. And you can follow the jump. But the thing is, that story, that front page story, was in the lower left-hand corner of the paper because on that very same day, human beings walked on the moon for the first time. And that previous path through the archive would not have shown you that. You wouldn't have had any context, unless you happened to be really good with dates; that January 20, 1969 was a momentous day in human affairs, and also a busy day in the news. It was also the day after Senator Kennedy was involved in a car wreck that has redounded throughout the years.

SANDHAUS: ***[10:47]*** The other thing that I really like about TimesMachine is that it lets you not only see the articles, but also the ads. And, the sort of pithy way that I like to talk about why this is cool is that the articles tell you what happened, but it's the ads that tell you how people were living. The Times happened to have several very cool ads in it here. This is one of my favorites, going to page 110, here. This is July 20, 1969. There's an advertisement in the paper for a music festival called "The Woodstock Music and Art Fair: An Aquarian Exposition," to be held three weeks in the future. It looks like a pretty good lineup. I think I might consider going to that. And, conveniently, all I have to do is clip out this ticket right here and mail it in with my check or money order, and I can get tickets mailed to me for Woodstock.

SANDHAUS: ***[11:49]*** The TimesMachine really has, I think, created a whole new realm of possibilities for our readers, and the general audience, to really delve into this important swath of American history. So, that's TimesMachine. It is something that we are very, very proud of, but it's an incomplete solution to the archive problem because, in the era of the web, users expect to be able to go to Google, and type in the name of an article, and get that article, and read that article like they would read any other web article. So, for that reason, we also launched an initiative for archive

transcription.

SANDHAUS: ***[12:32]*** The idea there is to take this old scan and actually pay an outsourcer to give us the full, digital text of those articles, and then put those articles online. This is our original review for "Star Wars" in both scanned microfilm form and in article form. And the big problem that we really wanted to solve with this is that most of our archive doesn't really have a good presence in Google. Sure, those abstract pages are indexed for a significant portion of the archive, but the absence of the full text and, even more important, probably, the fact that that text that was out there wasn't in our most modern web frameworks, made it difficult for Google to index that stuff.

SANDHAUS: ***[13:12]*** Well, I shouldn't just say Google. I mean it's search engines, aggregators, social media sites; all of that's important to have your stuff expressed in a legible way. So, the solution to this problem was, we ended up transcribing these archival scans. First, we transcribed 1964, and used that as a pilot, and decided, based on the results of that, to do 1960 to 1980. So far, we've transcribed all of these articles. We've published 1970 to 1979 online, and Google now indexes 672,500, approximately, new articles that were not previously searchable to the world. And now those readers can find the things they're looking for, come to us, and get that information in a way they weren't able to before.

SANDHAUS: ***[14:00]*** So, those are a few instances in which we, at The Times, have tried to take modern technology to give new life to things that were created with an older technology. But that's not all. We've also recently embarked on an era to modernize the web content itself, and bring that up to the standards and experience that one expects from a 21st century news publication. And to talk about that, I'm going to hand it over to Sophia.

VAN VALKENBURG: ***[14:46]*** So, what about our new articles? I'm going to talk about how we modernize our born-digital legacy web content. So, newyorktimes.com officially launched on January 22, 1996. And as you might imagine, the web was very different back then. But in 2014, newyorktimes.com got a redesign. The redesign included an updated article template, which better served the needs of a modern web experience. But there was only one catch: we could only serve articles using the new template if they were in our current content management system. What about all the articles published before we started using the CMS in 2006?

VAN VALKENBURG: ***[15:43]*** Well, they were stuck looking like this. So, our team's challenge was to convert our legacy web content before 2006 into a format that we could import into our CMS, so that users could view it in the new design. Essentially, our goal was to make an article from 2004, for example, look like it was published yesterday. Initially, we thought the problem was simple. We have an archive of XML files derived from print data, and we thought that all we needed to do was convert the XML into the JSON format required by the CMS. However, we quickly ran into an issue. We found that there were hundreds of thousands of articles from these years that were missing in our print archive. So, we inspected a sample list of web articles from 2004.

VAN VALKENBURG: ***[16:44]*** As it turned out, some of the missing articles did exist in our print archive, but we needed to combine the print and web data to properly process them. There were also some web-only articles that were totally missing from the print archive. So, we had to revise our strategy to include both print data and web data in our pipeline. Now, our questions became, what's the
complete list of articles from 1996 to 2006? How do we identify which of the missing web articles correspond to existing print articles so we can combine them and avoid duplicate content? Which articles are web only? How do we scrape that page for content and metadata? And finally, can we build a system that will process all this data for each year, easily and efficiently?

VAN VALKENBURG: *[17:38]* So, we started with a simple question: What do we include in our archives? It became apparent that there was no one definitive source of article data. We decided to compile the official list from four different sources: the print archive, site analytics from the past six months, movie, theater and restaurant reviews and also site maps.

VAN VALKENBURG: *[18:02]* Once we had our definitive list, we designed a data pipeline to combine and convert all of the web and print articles into JSON for the CMS. Here's an outline of the steps required for each year that we process. And don't worry, this slide will make sense soon. Here, the yellow boxes are inputs into the pipeline, the print archive and the definitive list of URLs. The blue boxes are intermediary steps, and the green boxes are outputs. The pipeline outputs a file with the articles in the new format, as well as any articles that failed along the way.

VAN VALKENBURG: *[18:44]* Using the definitive list of articles, we determine which articles are unaccounted for in the print archive. Next, we obtain the raw HTML of the unaccounted articles from the web. At this point, we still don't know how to combine the web and print articles, and we don't know which articles are web-only. In order to figure that out, we compare print article content to the web content using an algorithm that Jane will describe later. This allows us to combine the print and web data, remove any duplicate articles and determine which web articles are totally missing from the print archive.

VAN VALKENBURG: *[19:27]* Next, we reprocess the combined web and print articles and convert into the proper format for the CMS. Because we can use the structured XML data from the print archive, getting the article content and metadata was relatively simple. Then, we process the web-only data and convert that also into the proper format for the CMS. But, in contrast with the print articles, we only had unstructured HTML data for these articles. So, scraping the article content and metadata was harder.

VAN VALKENBURG: *[20:04]* Finally, we combine the outputs from the last step into one file and import it into the CMS. To give an example of the scope of this challenge, here are the approximate numbers for 2004 alone. Out of the approximately 116,000 articles for 2004, about 56,000 of them were fully accounted for in the print archive. Another 42,000 required data from both the print archive and the web. And 15,000 articles were totally missing from the print archive. Finally, there were about 3,000 articles that we couldn't process for some reason. For example, the article no longer existed on the web or it didn't have an article body. So, what we learned from this project is that it's important to use more than one data source when modernizing our archive in order to ensure completeness. Next, Jane is going to talk about some of the unexpected challenges we faced.

COTLER: *[21:09]* Hi. So, Evan gave us some background on the context for this project. Where were we coming from? What information did we have? What was the state of our world when we started this problem? And Sophia talked about the solution that we came up with, the broad pipeline that we designed to solve this problem. What I'm going to be talking about is all of the little things, some of them not so little, that we didn't anticipate and, in some cases, could not have anticipated; because they were problems that came up in quality assurance or problems that we realized were part of the scope of the problem that we hadn't known were there before.

COTLER: *[22:10]* So, the first thing that I'm going to talk about right now, briefly, is the first point: 1996. So, it's not just 1996. What that really means is the representation for all of the years in which we couldn't just run it through the pipeline because the year was some kind of hybrid, some data is

available and some data is available in a different way. For example, 1980 was like this. Evan mentioned it earlier, in 1964, since we'd run it through the pipeline previously and now it was slightly different. So, all of those years had to be taken into special consideration. The next three points I will go into after this, and then I'm going to follow it up by talking about the next steps, which are, what is the future of this project? What are potential problems that we're going to be facing after this?

COTLER: ***[23:10]*** The first thing I'm going to talk about is what Sophia mentioned, which was this step of the pipeline. How do we convert the print XML that we had with the web HTML that we scraped, as the list of what articles we had that were missing? We dubbed this process, affectionately, "fusion," because it was fusing one corpus of text with another. If you want to know, in detail, how fusion works, if you're a techie person or you're just curious, then you can look it up on our blog, The Times Open Blog. We wrote a post about it.

COTLER: ***[23:54]*** But to go over it briefly, right now, basically the essence of the algorithm was looking for common phrases between one corpus of text and the other. You can think of this by saying that some article might have a unique sentence that appeared in the print and also appeared in the web and is unlikely to be found in other articles from that year. So, once you go through and find articles that have a lot of matching words, you put it together and you say, okay, we think these articles match.

COTLER: ***[24:33]*** Another thing that we have to mention is that we tended to prioritize the print data over the web data, because, even though right now we see digital data as canonical, before the web was kind of an afterthought. So that's an interesting mindset to take into consideration with the advent of the digital era. The next thing that I'm going to talk about is search engine optimization. This is what an old article looked like before Sophia showed it. And this is the URL that it had.

COTLER: ***[25:17]*** So, it was "NYTimes.com/(year)(month)(day)/," in this case, "27IHT-scotus.team.html" which is completely unhelpful. In order to make articles easier to find in search engines, such as Google, and just to make them user friendly so that when you see a link, you know what it's going to go to, we changed the URLs to something like this, which is the headline separated by dashes, for example. As you can see, it's much more friendly. The problem with this is that if we just de-commissioned all of the old URLs, it would have led to link rot. I'm sure many sections on Wikipedia that link to many New York Times articles would have been very unhappy had we just de-commissioned all of the old, ugly URLs. So, as a result, we developed and currently maintain a system called Switchboard, which, if you can imagine what a switchboard is, you can unplug things and plug them back in.

COTLER: ***[26:28]*** It's essentially a redirect system. So, we store all of the references from the old URLs to the new ones, so that the other sites that point to The New York Times don't break. We really don't want that. There are more cases of missing data in this archive. Sophia already talked about the case of the missing articles and how we solved that with the fusion process. This is going to be the case of the missing sections. So, this is something that we caught in QA (quality assurance).

COTLER: ***[27:08]*** It explains that article transcription isn't perfect. You know, you can do OCR, and even if you do it manually, sometimes, in one percent of cases for example, there are going to be mistakes. This is just a section of an article that I screenshotted from the site. One section ends and then there's a section header and then another section starts. So, what we noticed during QA for some years was that some of the sections seemed to be blank.

COTLER: ***[27:45]*** We thought, well, this isn't good. Why isn't this transcribed? We can't put up

missing pieces of our archive. Well, it turns out, the sections actually weren't blank. It was just that, in the transcription process, there was a linking problem between the section headers and the text for that section. So, we made the user experience decision of, for those articles, taking out all section headings because we thought it's better to have all the text in the articles and maybe take out the sections than to have a very confusing looking article.

COTLER: ***[28:38]*** So, in that case, we decided to take out information that we had in favor of providing a better user experience to our readers. Finally, I'm going to be talking about next steps. The naive, first next step to do is to just transcribe more of our archive. We already have full text for all of the years after 1960, as well as all of the Civil War years, which is 1860 to 1865.

COTLER: ***[29:09]*** We hope to continue transcribing our archives so that, one day in the future, you can look at an article from 1851 on The New York Times website and read it as if it were published yesterday. We think that would be amazing and super cool. Lots of positive adjectives. So, something else that we want to do is put photos back into our archive. This is a screen shot from TimesMachine, from the Challenger explosion.

COTLER: ***[29:50]*** The New York Times is known for having many great photos, but unfortunately, in our current digitization of the archive, those photos are all missing because it was extra work putting them in, and we were really focusing on the text. So, in future iterations, we would love to be able to also publish the photos, and have them in the articles, and see how that would add ... well, it would add to the experience, but to see how we could do that.

COTLER: ***[30:27]*** Because, there's sort of complicated subtleties in, where were the photos placed, and how to refer to them, and all of that. Relatedly, and the final thing that I'm going to talk about in next steps, is we want to eventually figure out if there is a way to do digital preservation. What this means is, to figure out if there is some way that we can meaningfully communicate the original journalistic intent of our archive as it was in its digital form. This is not a particularly interesting thing to do that with because it has text, for example, but this would be more interesting to do with photos, for example, because there was a lot of editorial say in where do photos go and what photos are there, and all kinds of stuff like that. With that, I'm going to hand it back over to Evan to conclude.

SANDHAUS: ***[31:51]*** Thank you so much, Jane. Thank you, Sophia. The two of them have done, along with the whole team, a lot of really fantastic work. I should emphasize, on those previous slides, those are all things that we would like to do, but I don't want any of you to come away with the impression that those are things that we are firmly committed to doing next. There's an awful lot of stuff on the road map. I hope that we've gotten across the point, today, that modernizing and preserving the archives of The New York Times is an ongoing challenge and technology has this really annoying habit of not staying still. So, we're always trying to keep up with it. But, I hope we've given you some valuable context today, and I'm really looking forward, now, to the discussion. So, thank you very much.

MCCAIN: ***[32:38]*** Thanks, New York Times technical team. Obviously, a ton of work has gone into this, and it is fabulous to be able to go online and, with a few keystrokes, find articles. This is delightful, in that way. In my mind, if I'm thinking about an archive, I'm thinking about, "How did this really look when it was published?" And it seems like you've made the decision that that is not as high a priority as getting it into the modern Content Management System. Can you discuss a little bit about your process, your decision? I mean, there's obviously economics involved, but what else?

SANDHAUS: ***[33:31]*** Sure. I guess I should add the standard disclaimer that I am a technologist first and I am not a representative of our editorial staff. So, I don't want what I'm saying to be construed as the editorial position of The New York Times on these issues. But, it was my feeling, as somebody who was pretty close to this project, that the biggest service we could do to our readers, with our archives, was to put it into a format that made it easy to consume. But not just because the design is better, and the design is better, but also because, those legacy frameworks that the stories were published in weren't optimized for search, and probably even more importantly, weren't optimized for mobile.

SANDHAUS: ***[34:10]*** So, we thought it was a bigger priority to get these experiences ported to a stack that could support the sort of experiences most likely to make this most useful to the largest number of people. That, at least, was my thinking about why we approached it the way that we did. I do, however, acknowledge — and it's certainly become even more apparent to me over the last couple of days at this wonderful event that you've organized — that that is necessary, but not sufficient.

SANDHAUS: ***[34:57]*** And that we do need to think more critically about what can be done to preserve the original experience of these things, as was discussed in the panel earlier today. It's much easier to preserve print journalism because it's not this weird nexus of content and software. But, because these legacy templates really are an intersection of those two things, finding a way that we can preserve that, and in a way that we can preserve that for the long haul is, I think, a real challenge. Because I doubt anybody in this room has much faith that the current crop of web standards will meaningfully be interpreted by software in the distant future. So, figuring out how we can preserve this, without relying too much on the technology that was used to create it, is a real challenge.

COTLER: ***[35:53]*** I just wanted to add something. With what you said about us having the archive more in a modern format, it's kind of a decision to make it more easily accessible and less, I guess, faithful to the original. Even though we're going to be doing our best to put more of the original data in there. So, it's sort of like a lustful compression.

VAN VALKENBURG: ***[36:31]*** I just also wanted to add that another benefit of what we're doing is, as I had mentioned previously, we had all these different data sources. And the fact that we're standardizing the format that they're in will be very helpful, I hope, in the future, if we want to make further changes to the archive or do something new with the data; now we have it all in one format. So, it's easier to batch process, as opposed to having to go in and do something different for each chunk of years.

MCCAIN: ***[37:16]*** I've got to open it up in a minute, but access is clearly important. I mean, the news business is all about access. So, that part of it I get. Here's what's on my wish list, though. It sounds like you've taken some web-only content, and blended it with print content. Is there a way that I can sort out what appeared online-only, you know, that part of the experience versus the print experience, at this point?

COTLER: ***[37:51]*** Well, you can look at TimesMachine.

SANDHAUS: ***[37:58]*** I mean, it's an interesting idea, and it's not really a use case that we've given a great deal of thought to. It does involve a question that we've had to ponder, which is, "What is the definitive version of an article?" Now, traditionally the answer to that, at The New York Times, has been the Late City final edition of the article. Because that's the version that we send off to the

microfilmers, because it was the last edit of the story that made it into print.

SANDHAUS: ***[38:28]*** Now, of course that's probably not the right definition anymore. But, I'd argue that, over at least some of the archive that we've done the modernization work for, that was probably still a reasonable, operative definition. But, finding that right tradeoff between calling out that something was digital-only or print-only, I think, doing that well, would probably require us to do a degree of research into this digital publishing history of the organization, and really find out where that digital-only current became significant in the organization. That's probably where we want to start making those distinctions. But I totally understand what you're saying.

MCCAIN: ***[39:10]*** So is there a New York Times of record policy?

SANDHAUS: ***[39:16]*** Well, the official archival version of the paper has traditionally been defined as the Late City final edition. I feel a little silly admitting this, but I'm not totally certain what the official definition is, at this point. But that has been the traditional definition.

MCCAIN: ***[39:35]*** Questions?

AUDIENCE MEMBER 1: ***[39:46]*** Thank you for your presentations. If, say, a researcher wanted to do text analysis on the entire corpora of what you have in your collection, would they be able to do that? Would they have to have a subscription? What if they were working in a group? Could they get one subscription for the whole group? Would they be able to do text analysis on the entire corpora?

SANDHAUS: ***[40:14]*** That's an excellent question. And, one that we've tried to answer by something that we did several years ago now, that I was involved in, called The New York Times Annotated Corpus. I'm sure many of you were familiar with this organization in Philadelphia, called the Linguistic Data Consortium, which is kind of like the chemical supply store for information science and computer science experiments on language data. The Linguistic Data Consortium has this collection, The New York Times Annotated Corpus, which is available to members for free and to nonmembers for a few hundred dollars, that contains every article we published between 1987 and 2007.

SANDHAUS: ***[40:58]*** Now, I realize that's not the entire sweep of the archive, but it's a fairly substantial collection. It's well in excess of a million articles and a billion words, and a dataset that, last time I checked Google Scholar, had been the subject of somewhere in the neighborhood of 300 scholarly articles. So, we're really happy that that data is out there and that it's available to the community. Now, if you wanted to do something more broad, the answer is, well, you could probably double the amount of data, but pretty quickly you would run out of full-text data. We don't really have that going back beyond 1960. And, if one were interested in working with that, I think the best strategy is to get in touch with me and we can see what sort of arrangement we might be able to make.

AUDIENCE MEMBER 2: ***[41:46]*** Thank you for a very interesting presentation. I enjoyed it quite a bit. My question has to do with something that's often left behind in projects like this, and I wondered if it was a part of your project. You've obviously made a lot of decisions along the way to go one direction or another. I wondered if there was a documentation process for this whole thing, above and beyond something like this. But, actually, are you documenting all the different decisions, which is kind of a preservation type of activity?

VAN VALKENBURG: I will say, as was mentioned in an earlier talk, GitHub is a version control

system. All of our code that we've been using for this project, we've been storing in GitHub.

VAN VALKENBURG: ***[42:41]*** Pretty much, all of the — if you use version control as a type of archive, all of the changes that we've made along the way will be preserved there. So that's one way that we could be able to go back and see what decisions we made in code.

COTLER: ***[43:07]*** Also, in terms of ... well, Sophia answered about code, I'm going to talk more about design decisions. In terms of those, we have regularly scheduled meetings for this project, and so every week we record minutes of what decisions we made. So, there is a document that we have that details the history of all of the decisions we've made. Is it in a super organized form? No. Because we didn't really put it into that. But, we do have a record of all those decisions, yes.

SANDHAUS: ***[43:42]*** And this is not the first contact migration The Times has done in its history. I can assure you that one of our goals is to make future generations happier with the documentation around this migration than we've been with previous migrations.

AUDIENCE MEMBER 3: ***[44:02]*** I have two questions. Jane, in particular, mentioned user interface decisions. How did you arrive at those decisions? AB testing, parking users into a room and watching what they do, or some other way? And the second question is, how have you future-proofed your current design for storing data?

COTLER: ***[44:31]*** So, I guess I'll take the first question. Well, in the example that I mentioned with the appearing to be missing sections, what it really was is that, when we were QA-ing it, we thought, "This is very confusing." So, we tested it on ourselves, and we kind of sent it around to people, internally, to make that decision. I can't say we did AB testing, but we did come to a general, and pretty wide, consensus that, in that case, that was the better option.

VAN VALKENBURG: ***[45:14]*** Also, I just want to clarify that our team is actually not involved in the actual front-end design of the web pages. So, that's an entirely different team, which is focused on that. And they have put a lot of work into figuring out what is the best user interface design for, not just this app, but all of our web content.

SANDHAUS: ***[45:49]*** And, one of the things about the design that I'm really happy about is that ... So, we haven't migrated all of the content. Right now, we've migrated all of our content from 1996 onwards and from, basically, 1970 to 1979. So, there's still sort of a donut hole in the middle of stuff that we need to migrate. But, the decision that we made in collaboration with our design folks, that I'm really happy about, is that we decided that ... we launched — well, we didn't decide that — we launched our website in April-ish of 1996. Our website just turned 20. There were blog posts about it.

SANDHAUS: ***[46:29]*** And we made the decision that, in the design for the archive pages prior to that day, in 1996, we link to TimesMachine. Because, prior to that day there wasn't a website. So, we've just made the decision that, prior to that day, it's important that that image of the newspaper and that link to that print experience, we felt that was important to provide with our users. And we were gratified that the design department helped us realize that vision.

VAN VALKENBURG: ***[46:59]*** You'd also asked about future proofing, and I'm going to go back to a point that I made earlier about the fact that we've combined all these disparate data sources into one

format. So that's, I think, one way of future proofing. Now that it's all consistent that's a lot easier to deal with when we want to make changes in the future than dealing with different sources. So, yeah, I'm sure, you know, hopefully if The Times is around for another hundred years that there's going to be more changes, but it'll all be in one format now.

SANDHAUS: ***[47:43]*** But I think, ultimately, when it comes to future-proofing this stuff, if anybody here knows what the future is, I want to talk to you. But I think you just do your best and hope that the decisions you're making are going to lead to a good outcome.

AUDIENCE MEMBER 4: Is all of this stuff both mobile and desktop friendly?

SANDHAUS: All of this stuff is both mobile and desktop friendly.

[Audience question inaudible].

SANDHAUS: Sure. Right now, this content is available almost exclusively in English. I should say that our coverage today is available in, at least parts of that coverage are available now in three languages. We publish a Chinese language website. We publish a NYT en Espanol. So those three languages are the languages primarily served by our organization. But in terms of the archive stuff, that stuff is all in the original language of publication, which was English.

MCCAIN: ***[48:47]*** I think that's going to be the last question. Thank you, New York Times technical team.