Technology and Platforms: What's on the Horizon

Georgios Papadopoulos
Atypon, georgios@atypon.com

Georgios Papadopoulos, "Technology and Platforms: What's on the Horizon" (2017). *Proceedings of the Charleston Library Conference.*
http://dx.doi.org/10.5703/1288284316709

Technology and Platforms: What's on the Horizon

Presented by Georgios Papadopolous, Atypon

The following is a transcription of a live presentation at the 2017 Charleston Conference.

Georgios Papadopolous: Thank you very much. Thank you everyone. Good morning. Thank you for the invitation. It's a large crowd, so I'm the nerd in this, so of course I'm going to talk about technology. For 22 years, I have been running Atypon, and Atypon started as a dream for creating a better technology company for scholarly communications. What we perceived as not a very strong technology from the various players that existed at the time, we had a very strong focus on technology and enabling publishers to do more things with their websites. Atypon has grown tremendously in the years since it was founded. It's hosting about 40% of all scholarly research content right now. So, you are interacting with some of our websites, and the reason for my presentation today is I'm going around trying to incite change on both the publisher side but also on the library side. We're sitting in a place as a technology company, as a technology service company to the publishers, we're sitting at a place where we develop a lot of technology but we cannot necessarily launch it until the publishers demand it. And in some cases, the publishers also want to hear from the librarians that the librarians are ready to embrace this change, so there is an ecosystem and there is a whole community that needs to embrace some of these changes that I want to talk about today.

I've been there since the beginning. I've been there since the first journal that launched and I was responsible for it, and the first two years were very exciting. We were doing a lot of different things, that was with Highwire at the time. If somebody, however, would go into a long sleep in '97 and wake up November 8 of 2017, and if you would go and look at the websites, he would see pretty much the same things that he saw before he went to sleep. And what he would see is basically, for institutional access, you would have IP authentication pretty much, username and password for your individual accounts as a user, XML as the format that the content is coming in, a very dumb form of HTML and PDF, which are both pretty much dead if you look at them for a reading experience. The big search engines for if you have any questions that you want to ask across different publisher sites, and of course e-mail alerts that you have to register on every publisher site if you want to get any alerts about what is new, and frankly, even in '97, there was some archiving going on and the same incomplete forms of archiving that existed then pretty much exist now. There is not too much change. What is interesting in technology for 20 years we've been doing pretty much the same thing over and over, and I can tell you Atypon is like 350 people, most of them engineers, and we're doing pretty much the same thing and the same thing happens all around the industry.

Why do we need change? Maybe what we're doing actually works. Who said that we should be doing any change? Well, I tried to list a number of reasons. I can list 100 more reasons, frankly. But some of the reasons, some of the topics that I've put here and some of the questions that I've put here and reasons that I put here touch on some of the topics that I want to talk about: institution authentication and PDF drive content piracy. SciHub, LibGen, and all of the other forms, and you might have your own views on piracy, I think that it actually has helped technology companies as well but at the end of the day it puts, it actually threatens the fundamentals of scholarly communications. Without rules, there is no game.

So, then of course we have user frustration with the various authentications that they have to do over various websites. Personalization is good. You know, everybody should have, every site should have personalization. The problem is that it becomes a nightmare for the users. And the, I call them archaic formats, HTML and PDF, basically is what we had 22 years ago, 20 years ago, basically they just mimic the print. That's what we do. We don't do anything much about it, although in the current world we see, we want more digital interaction. These formats don't support digital interaction at all. Search engines, discovery done through search engines and search engines are good if you know what you're looking for. They're not good if you just want to manage the new information coming in. Getting 100 e-mails per week is no way, is not so manageable. And of course, as I say, archiving that is increasingly missing more and more content because more and more content in our websites, I can tell you, is not the content that

https://doi.org/10.5703/1288284316709

it used to be 20 years ago; 20 years ago, we were just receiving the articles. Now you're receiving all of this other content around the articles, editorial content, that is actually not archived anywhere. So, all of this is actually lost. At least it is kept by the website somehow, but a lot of it is actually going to be lost eventually. So we need to do something about it. Okay? Then there are other things like annotations that the users are putting in, and that is also going to be lost because that is also content in many cases, so we're increasingly seeing user-generated content, editorial content, all kinds of content that is actually not archived anywhere.

Let me talk a little bit about the changes that are coming and how they are actually related to the libraries. There is a big initiative called RA21. It is a joint STM and NISO pitched initiative with publishers, technology companies, and libraries, as a matter of fact, that are trying to change the way the library patrons access content that they have access to. We've known the problems; we know that it is not convenient. We know that it is not—it actually enables content piracy and we can really have the Holy Grail here. We really can and we've proven it with some of the prototypes that are going on that you can have both seamless, very convenient access, preserve actually individual privacy, which is very important, and prevent content piracy. So, this is something that needs to be embraced by the libraries as well because there are some changes that need to happen. And we think that this is going to be rolling out in 2018 and 2019 so, you know, you get some information on this one. Now, it's very interesting to me, the other one is very interesting, the individual user authentication. There's so many social, academic social networks, any one of these could become essentially the SSO for users to access and have a password only in one place. None of them are vying for this position, interestingly enough. Nobody wants to do it. It's a problem that needs to be solved; any of them can solve it. I would invite them to solve it. That's something that maybe you should press on.

Content formats. I call them dinosaurs. Digital in name only. So, it's a bunch of compromises the way they are today. You either have something that is portable, like the PDF, and immersive or you have something that is actually a little more dynamic and has a little more interaction, that is HTML. You either have something that is adapting to the device like the HTML or if it's PDF it is very hard to read on the phone or you have the PDF basically. The data right now is all linearized into pictures, which is what we were doing for print, so you will not perhaps understand the difference unless you are shown the actual data, but when we're printing articles the researchers had to take the results, create a chart, take an image of that chart, and send it with their submission. We do exactly the same thing now. Although clearly the researcher could just give the data. They could just say this is the way to create a chart out of the data and the chart can actually be created right there, and if I want to change the chart because I want to see a different kind of chart, I can as a user. That's what I want to achieve. The data usually is stored somewhere else. It's somebody else's problem, these linkages, who knows whether they are going to be preserved or the data is going to be preserved? Most likely that is also going to be lost.

Just to give you, I stole this reference, actually, from the Scholarly HTML site, I think it's attributed to Sebastian Ballesteros. Basically, *New York Times* had a report that did some change to their websites and they just added structured data. And by just adding structured data it increased our traffic to our recipes from search engines by 52%. So, as Ballesteros says, in other words cupcake recipes are reaping greater benefits from modern data practices than the whole scientific endeavor. And it's true. It's true. I mean it's sad, but it is true. This is the state of the art in scholarly communications, which is insane because a lot of us are PhDs and we really care about the subject, we are really into it, not only as people who are trying to do things, but also as leaders, and for some reason this has become tough, difficult, so part of the difficulty is the freedom of what you do for the formats? Think about it, science has the richest content in terms of information. It has the richest information. We can really have a very rich information fabric on articles, yet nobody seems to be doing it.

Where do we need to move? We need to move to Scholarly HTML, and I'm not talking about moving toward Scholarly HTML just for reading. It's going to become the format, I believe, for authoring content as well because we really need to take the view that the content is going through several phases and it's enriched in several phases and we really need to keep it together. All of this content being created in all formats transformed into XML, losing a lot of the original, then trying to re-create from XML some delivery format and losing a lot of what's going on in between is just nuts. So, we're moving toward Scholarly HTML, and the beauty of it is it is immersive if you do it in the right way, it adapts with the device, it allows annotations that is user-generated action, user-generated content, it invites that. We need to make it portable, and to make it portable it means that we need to use ePUB, a very nice standard. So, essentially what I see is that in the next year or two, and this is what we are working very actively on, everything goes into ePUB with HTML. Data remains in the document, okay, and is something that you can repurpose. The user can actually view, for example, as I said the results of the experiments with different viewers, in different ways, in the ways that make sense for him or download it right there. Semantic overlays will allow extensions to the paper and add comprehension to it, and of course machine learning, machine learning is the future of scholarly publishing as we're trying to extract more and newer information and as we understand more ways in which we can extract the rich information that is in the papers.

Discovery. It's good to have all of this information. How good is it if we cannot actually find it, or if there's so much of it that we cannot actually get to it? Again, search engines are very good, excellent when you have something, a specific question, and you have a way in which you can approach it, then you can find somehow your information or parts of your information. However, researchers, what they need is actually to know is what is new? It's the fear of missing out, they need to know what is new daily, and if you are a biomedical researcher this is becoming just, as you know, impossible. There's probably, you'd have to read around 400 abstracts a day if you want to keep up with any slice of a domain. So that's becoming simply impossible. Getting e-mails from journals sent to you is also not a good way of managing your information, this firehose of information aimed at you.

So, robots. Paraphrasing a famous movie, one word: robots. Robots will help us. So, yes, I know librarians will also help us, but with thousands of papers a day, it's becoming very difficult. I think the librarians will help us, will help users in purposing their robots so that they can actually enhance the discovery. So what you will see is personalized tools that know the user and they will bring from the websites what is important for him. So, and they will take into consideration everything that he is reading and his social networks, all of the information about your social networks.

Going a little faster, archiving, current problems with archiving, as I said most archives, there are two kinds of archives. There are archives that capture just the content and they capture it in XML and PDF form, and then there are archives that capture the site. The problems with the XML/ PDF, the static archiving sites, is that they are increasingly missing out content and missing out user-generated content that is attached to the actual content. For example, annotations and comments and so on and so forth and all discussions and the site archiving sites, well, they don't know what they're capturing as a matter of fact, because they don't know what changed, when it changed, or anything like that. So you have something like the Wayback Machine, where I said just 10 years ago go back to the Atypon site and let's look at some of the pages, and this is one of the pages that comes up. So, something that they did capture.

Archiving can actually be perfect. And I mean perfect. You could actually make it, if you remember there was a problem some years ago that even if we have the archives we're not going to be able read them because the tools that we used 10 or 20 or 50 years ago would not be possible or would not be even available to install anywhere. That's no longer true. Technology has solved that problem. What we need to solve is the problem of archiving our websites, and the only way to archive the websites is actually from the CMS itself. I haven't figured out the governance, what the governance needs to be. Technically it is a difficult problem. I'm not going to say that it is not a difficult problem or that it is an easy problem. But, the CMS itself, the content management system itself, knows what changed, when it changed, and it can do perfect archiving. You can go to any date and you will be able to see the website and interact with the website the way that the user of that website would interact in that day.

In conclusion, basically technology has stayed the same for 20 years. Let's make a pact that every 20 years we will be changing so that it will keep our lives a little bit interesting, and what we will need is also the help of librarians to push sometimes the publishers to make some of these changes. Thank you very much.