
**Interview with Toby Green, Coherent Digital**

**For podcast release**
**Monday, November 16, 2020**

KENNEALLY: According to legend, fire devastated the Great Library of Alexandria during the siege of the Egyptian city by Julius Caesar in 44 BCE. Tens of thousands of papyrus scrolls were said to have burned, creating a caesura in the corpus of ancient Greek literature, and the loss to humankind of countless artistic and scientific treasures. In our own time, when research and analysis is found increasing in digital form, a similar break in the knowledge timeline has started to open.

Welcome to Copyright Clearance Center's podcast series, I'm Christopher Kenneally for Beyond the Book. The recently launched Policy Commons hopes to fireproof the online library of the Internet, at least where it comes to policy, documents, and research papers. Policy Commons makes available for discovery and access nearly 2.5 million documents from thousands of IGOs, NGOs, research centers and think tanks. Policy Commons is the first project for the startup Coherent Digital, which is collaborating librarians, technologists, publishers and academics to tame large bodies of content and make valuable information cohesive, understandable, harmonious, and coherent. Toby Green is company cofounder, and he joins me now from just outside Paris, France. Welcome to Beyond the Book, Toby Green.

GREEN: Well, thank you very much and that was one hell of an introduction. Thank you.

KENNEALLY: Well, I always love a bit of context, and I think there is absolutely a parallel with the Library of Alexandria in the sense that we take for granted that the store of knowledge will be with us forever, and what you are trying to solve is this problem that it is not always with us forever, and, in fact, even if it's there, we can't get to it.

But let's just give people some background on your work. You have a long track record of innovation in digital publishing, Toby. You were Chief Operating Officer of Public Affairs and Communications for the Paris-based OECD, the Organization for Economic Cooperation and Development. OECD researches global development issues and provides policy guidance for its member nations. I

wonder if NGOs like OECD are concerned that their work is getting lost in the ocean of information that is the Internet?

GREEN: Well, in a way, I would say, I wish they were concerned because the reality is that international organizations like the OECD and the World Bank, and the IMF and the UN family, or non-governmental organizations like Chatham House and Brookings and so on, their main focus is on doing their research and getting the findings of their research and their work out to their stakeholders and, they hope, to a broader public. But they tend not to employ people with publishing skills, and as a result, they push their content out on their websites and they don't go the next step, which is to add that wrapper of metadata around the content that happens in scholarly publishing, and therefore makes the content fragile, it makes it at risk of disappearing. I think that's the big difference between the thousands of policy organizations and scholarly publishing, in that the former don't use publishing techniques that the latter do. That gap really matters because it makes the content fragile, it makes it at risk of disappearing, quite apart from the fact it makes it very hard to find.

KENNEALLY: That's a very helpful point you're making there, because I thought at first this was a preservation challenge, but you make it sound more like an organizational challenge.

GREEN: To go back a bit to my days before I joined the OECD, I worked in scholarly publishing, I worked for Academic Press and Pergamon Press, and then Elsevier. So my entire background and early training was in scholarly publishing, and I was very lucky to be involved in some of the very early experiments to put content online or even on CD-ROM. I've always been really intrigued by the challenge of how do you, in a sense, get the content from the author to the reader? And scholarly publishers, together with libraries, they've built a whole ecosystem that makes that work.

So when I joined the OECD, I was actually quite shocked when I got there to find what was in effect a mid-size publishing operation, they put out about 300 books a year. But they did so in their own peculiar manner, by the only standardized system that they used that was in common with the publishing industry, were ISBNs. But everything else, they did themselves. They even did their own printing. They didn't use any of the standard systems. Their content wasn't discoverable in the main discovery channels.

Curiously, when I visited libraries and introduced myself, and the librarian would say oh, yes, we've got all your books, come over here, and they would actually lead

me usually to a room at the back of the library which would be full of all of what they called the official documents. And there, organized in alphabetical order, you would find all of the World Bank content, all the IMF content, the OEC contents, and so on.

I would end up teasing the librarians by saying OK, great, well, I can see all my publications, but can you show me where all of the Academic Press publications are, or show me where all the Wylie publications are, and they would go, well, no, but they're organized by subject. And I said, well, why haven't you got the OECD's education books with all of your other education books? Oh, because you're different. I realized that there was this treatment of governmental official and non-governmental organization that's treated differently to scholarly publishing, it's managed differently. Therefore as a result, it was ignored, and the usage of the content was incredibly low.

So what I set out to do at the OECD was to basically normalize the way that the OECD did its publishing, and to use the tools and the techniques used by scholarly publishing, and to push the OECD to make sure they were using the same tools and techniques. So we started using DOIs, and we started getting our metadata into the main discovery engines and so on, so forth. We built something called the iLibrary, which is a platform that behaves just like any other scholarly publishers platform, and you can cite the works in the same way, and the things that are compatible with the citation tools and so on.

The reaction to this was – well, two things happened. One is the usage of the OECD's work went through the roof. We increased reading at least 40-fold in the time I was there. But the other thing that happened was the librarians and users were saying, well, this is fantastic, we can now find the OECD stuff. Could you now do the same with other international organization content, and in particular with non-governmental organization content? Because we know it exists, but we can't find it. Or if we do have a link to it, we find that link is broken. And they wanted basically an iLibrary-like platform to house and pull together all of this policy content from these thousands of organizations.

Now, at the OECD I looked in to see whether we could, in fact, do that at the OECD, but the OECD isn't a government organization. By law it's actually not allowed to offer services to non-treaty based organizations. So you can work with other UN bodies and so on, but what it can't do is to work with NGOs. It can make a publishing platform for NGOs.

So when I left the OECD just over a year ago, I got together with Stephen Rhind-Tutt, and we created Coherent Digital, partly with an eye on solving this problem, to basically build an iLibrary, if you like, for the NGO/IGO/think tank content because we know that this content is sitting out there, we know it's fragile, we know it disappears, links break. When funding runs out, particularly from a smaller think tank or a smaller NGO – when that funding runs out, that website is just simply switched off. There is no locks or clocks or portico for this content. Worst of all, this content just isn't found. It's really hard to find.

A friend of mine runs an NGO out in California, and I know him because he and I played frisbee together years ago, and I challenged him to say, well, look, Scott, (sp?) I can't even find publications on your website, so you do have them, don't you? And he said, yeah, and he gave me the link, and sure enough, they'd got a little repository. And then he said, it's funny, no one ever seems to find our publications, and it's because they just sit in a tiny website in a corner. They're very small, they only put out about 20 reports a year. Despite all of their best efforts through social media and so on, they're only ever going to reach a certain number of users. But if someone goes to a library discovery system, or someone even goes to Google Scholar or anything like that, they're simply not going to find this content.

So part of what we're trying to do is not just make it safe, we actually want to boost the usage of the content, make it more discoverable so people can actually access this knowledge.

KENNEALLY: And the way you refer to this content, Toby, is as wild content. In other words, in contrast to this tamed or domesticated content that can be fairly easily found in libraries. And this concern for taming this content isn't just a professional matter, but it really does have an impact on our lives. You make the case that there has been a tremendous amount of research done on pandemics in the past, and that research has been difficult to access at this moment when we need it the most.

GREEN: Quite. Last year in October there was an event which modelled a coronavirus pandemic in NY – I mean the event was in New York. It modelled a coronavirus. It was well supported, it was organized by the Johns Hopkins University, who are now famous for their coronavirus data with the World Economic Forum, and backed by the Bill and Melinda Gates Foundation. They got together a team of experts, and they spent a whole day modelling what would happen. The results of that meeting are on their website. That's it. And I challenged libraries to say, OK, well, how many of you have got that content in your library systems? Who's archiving that content in the long run? When that website gets switched off or it

gets corrupted or something, who's going to look after that content? Well no one is. Within a fortnight of that event, two NGOs put out reports about the state of the world's preparedness for pandemics. This content was published last year. It's there, it's freely available, but no one knows how to access it because it just doesn't appear in discovery systems. And it can disappear.

One of the major UK think tanks rebuilt their website a couple of months ago. All the links to their content were broken. If you had a bid guide or you had a syllabus pointing to content on that website prior to two months ago, that link will be broken because when they rebuilt their website, every single link broke. This is a major think tank in the UK.

In building Policy Commons, we've identified over 200 think tanks and NGOs that have gone out of business, and we've been tracking down their content and recuperating it, and making it available again inside Policy Commons. We found someone attempting to build a policy commons back 10 years ago, they ran out of funding after they got to about 30,000 records, and so we've now saved that content. All that content's now available inside Policy Commons.

I'm in discussions at the moment with an African archive that's pulling together NGO content in the continent of Africa, and they've got about 5,000 reports in their repository. Their funding has run out – it runs out at the end of this year – and if they don't get new funding, that service is going to come to an end. So again I'm talking to them to see whether we can help keep what they're doing going.

And this content is valuable content. It's just as valuable as the content is in journals or book series from the major publishers.

KENNEALLY: Right. And Toby Green, this challenge that you're describing in pulling together Policy Commons – well, actually there were two sides to it. There's a technology challenge, but there's also what I guess, looking at it from the perspective of Copyright Clearance Center, would be a permissions challenge, a rights challenge. Describe how you approach that latter challenge.

GREEN: The copyright regime used in all these organizations is incredibly varied, that's the first thing to say. Some just use what I would call traditional copyright, some use the Creative Commons suite. Some, if you look on their websites, and look inside their publications, they make no copyright statement whatsoever. So we've got the whole spectrum of challenges there.

Now plainly, if we were to try to write to them all to say, hey, can we have permission to put your content inside an aggregation service, it would require a huge amount of effort from our side to chase them down because in a lot of cases, they simply won't reply. So what we decided to do instead is that we're culling their websites, we're extracting the metadata and the full text of the items, and we use that to build a discovery engine, which you will find inside Policy Commons. So within Policy Commons, you search, you find the content that you're looking for, and then we give a link back to the original website. That means that the full text is retrieved from the original website. Now that gets us around the copyright problem, but I also think it's really useful for the user because it means that the user gets to see the content in its original context, and there is a value to that. It also means we're driving traffic back to, quite often, these small, nichey (sic) organizations, and that gives them the additional traffic which could be very useful for them in terms of their ability to get future funding.

So we see this as a very positive thing to push the traffic back to them. However, we've got a copy of their report, we use that to fuel our full-tech search engine. If that link breaks, what we will do to our members, so it's our people who are paying us, is we will give them access to that saved copy. Now, we're reaching out to organizations to say that this is what we intend to do. If you object, obviously we will respect that, and we won't offer that content out, but our default is that we will offer this service. From our point of view, we see it as an insurance service against link rot. We don't lay any claim owning that content, and we make it very plain to our members that if they were to access and download content from that saved link, it is on the basis of we've given you a replacement link. And if the copyright owner comes after us and asks us to take it down, we will have to respect that.

KENNEALLY: Toby Green, let's go back to that image I opened with, the Library of Alexandria, and the purported tragedy of that fire. You described the concerns you have about the research related to pandemics and the coronavirus, but it's not limited to matters of health policy or public health. The kind of issues, the topics that you are working with, I'm sure, are quite comprehensive, just every manner of subject. But give us an idea. Something like thousands of subjects are covered, and topics are covered in Policy Commons. Just give us a quick sampling of the sorts of things we could find there.

Q: Well, we 7,300 topics inside Policy Commons, so we really are covering every aspect of policy, which basically covers every part of human life. My brother-in-law is an architect, and he was having a look at Policy Commons. I thought, well, you're not going to find anything of interest. No, he put in the word architecture, and he retrieved a huge amount – 10,000 hits he got. We're finding stuff on

religious – policy around religion. We've got policy, obviously, the big issues like Brexit, trade relations, economics. But it's really quite extraordinary what we're finding content on. I think that, certainly, the feedback we're getting from users is that they're constantly being surprised by wow, I had no idea that this content existed.

There's a researcher who we've been using and working with a health development search engine, but she's an expert in her own particular field. When we rebooted the search engine and asked her to have another go at it, and she said, look, I've just written a report, I've just been doing a great, big study in a field, I thought I knew everything. I found new reports inside Policy Commons that I had no idea existed. That to me is proof that we cans surface content that otherwise is really hard to find. She actually measured us against. She did a search on Google, and then she did the same search in Policy Commons, and she said, I'm never going to start my research using Google again. So I thought that was really fantastic.

KENNEALLY: Imagine a world where you can survive without Google. That's quite an achievement. Toby Green, cofounder of Coherent Digital, which has just launched the new Policy Commons. Thank you so much for joining me today on Beyond the Book.

GREEN: Thank you very much for inviting me.

KENNEALLY: Beyond the Book is produced by Copyright Clearance Center. Our co-producer and recording engineer is Jeremy Brieske of Burst Marketing. Subscribe to the program wherever you go for podcasts, and follow us on Twitter and Facebook. The complete Beyond the Book podcast archive is available at beyondthebook.com. I'm Christopher Kenneally. Thanks for listening and join us again soon on CCC's Beyond the Book.

END OF FILE