GOOGLE BOOK SEARCH: OPPORTUNITIES FOR ACADEMIC PUBLISHERS AND AUTHORS IN THE ONLINE CHANNEL

Presented by the Text and Academic Authors Association

M: And I want to introduce Chris Palma by way of his biography, and he'll be talking to us about "Google Book Search: Opportunities for Academic Publishers and Authors in the Online Channel."

Chris is a strategic partner, development manager for Google Book Search. He's responsible for developing partnerships with book publishers of all types in the Western U.S., Australia, New Zealand, and Singapore. I think you just said you came back from Australia.

PALMA: Just got back from Melbourne.

M: Chris spent 15 years in scholarly publishing with Yale University Press and Harvard University Press, leaving his sales director position at Harvard in 2000. He briefly joined a DRM technology startup company, Content Guard, then moved to Palo Alto-based ebrary, a leading provider of e-books to the institutional library market, where he was vice president of content and business development before joining Google. Chris.

PALMA: Thanks very much.

(applause)

PALMA: What I thought I would do is give you a little bit of kind of background on how I got here, because I think it's an important story because it kind of mirrors some pretty revolutionary changes in the academic publishing world.

I was mentioning to someone earlier, when I started in publishing at Yale University Press in 1985, you could publish any monograph on 16 seconds in the Renaissance and sell 1500 copies. The vast majority of those went to top research libraries and assorted other institutions, and maybe about 20% would be sold directly to academics. When I left publishing in 2000, that number had dwindled down to about 300 or 400 copies.

Meanwhile, a big chunk revenue had been taken out of the equation for university presses due to the inflated costs of scientific journals. So librarians really had to make a call here, right? They looked at the circulation data eventually and they suddenly realized that these books weren't leaving the shelf. A very frightening number from a study back then said that 50% of all humanities monographs circulate .5 times in their entire lifetime on the shelf. So is it no wonder that librarians had to cut, they would cut their book budgets. Now, is that because there were not 1500 potential buyers for that monograph? Probably not. In fact, I'm almost sure it's not the case.

So for presses, we were being squeezed on a sort of pretty dependable revenue side for us, the monograph side, which is three-quarters – for some – three-quarters of their publishing revenue. So at least at Harvard, there was one recourse. We invested more in trade publishing, which of course is quite a bit more speculative, and we had some very nice successes. We published a number of bestsellers and near bestsellers, people like Eudora Welty and Toni Morrison and Edward O. Wilson and others.

However, the trade houses were looking over the fence and saying, hey, these guys are doing a great job. They're actually selling some books out there. So you'd have one shot with that author and then the next thing you know, that author would be pilfered by some trade house who would offer a monumental advance, only for that book to probably go out of print in five or six years. But that was the economics of it anyway.

So it was very clear that something had to happen, and what did happen was the Internet emerged as a revolutionary new media form about the time that I was starting to get pretty disillusioned with our ability to meet our mandate, which was to disseminate scholarship on books and projects that no commercial house would see as viable.

So around about that time, I decided to leave a very cushy job at Harvard where I'd been for 15 years to leap into the high-tech world about six months before the bubble burst. And that's typical of my impeccable timing. But I've learned quite a bit since then, and what I want to talk about to some extent is these changes in the book business overall and these sort of revolutionary forms of adoption in terms of the Web.

What I thought I'd do is, I'll start off just talking about kind of Google and search and what it is and what it means and what we represent out there in the world, what our business model is. A lot of people kind of don't know. Are we friend, are we foe? Where do we fit in the world? I'll pause. I'll take some questions for that segment, then we'll talk about the book business and then finally, we'll drill down into Book Search. So feel free to stop me, interrupt me, at any time through this because we've got plenty of time.

This slide is from two months ago and it is clearly out of date already, just to give you an idea how fast Google's growing. I started in September of 2006 and I think I was employee number 6008. Something like 6008, and we're now up over 12,000 employees. We're in 40 countries. We represent – right around 65% of all the world's searches are done by Google, if you looked at the top line. In the United States, we're maybe 49% to 51% of all search. Places like China, the second largest Web market, we're only about 20%, so we've got some work to do there. But if you looked across the board, we're about 65% of the marketplace.

If you looked at all our products – so you looked at G-Mail and Calendar and Search and Book Search and Scholar, and you looked at our Ad Sense advertising programs, which we syndicate through thousands of partners, if you look at that as a whole, we're reaching, every month, 80% of the World Wide Web audience, which is pretty significant. I don't spout these numbers to sort of puff out my chest and try to make a case for how grand we are. I point them out because there's a huge opportunity here for publishers, and there's a huge opportunity to disseminate knowledge and to democratize scholarship in a way that's never been possible before. And that's why I'm here.

For those of you who were here last year, I'm sure my colleague Rolin Lange (sp?) spouted the company mission statement, which hasn't changed since the two founders ran Google out of a garage in Menlo Park, and that is to organize the world's information and make it universally accessible and useful.

So around about 1999, 2000, Google had indexed a billion Web pages and it proudly stated that right on its clean Web page. It said, a billion pages indexed, like McDonald's, a billion served, right? It quickly became absolutely ludicrous to even talk about a billion pages. The Web is growing at millions of pages a day. We've got the introduction of wikis and blogs and social networking sites. The number of pages being authored is just phenomenal. Every new index refresh finds 10% to 20% of the Web that's completely new, just to give you an idea of how fast things are changing on the Web.

However, right about that time, Larry Page and Sergey Brin, the founders, realized that if we really wanted to commit to this very ambitious statement, we had to find a way to get off-line content into the index, because if you're going to talk about relevant information, it doesn't get much more relevant than book-based information and journal-based information. And we also extend into news archives and video and a whole host of other previously off-line-based content. So it was quickly realized that we had to find a way, and we had the resources to do it, to get this content up and live, and Google Book Search was born in 2004.

The quote here from our fearless leaders is, the ultimate search engine would understand exactly what you mean and give you back exactly what you want. So if this is getting to that ultimate search engine, Google, as good as we are, right here,

because the vast majority of people aren't good searchers. They don't enter multiple search terms, they don't us Boolean logic. It's up to the sort of science of intentionality, which is what Google faces every day, to determine exactly what people are searching for and actually give them the best results. The reason why many of Google's products, including Book Search and Scholar, are in beta for many years is that the way search science works is you write the best algorithm – and these guys are the smartest guys in the field writing the best algorithms for search – and then you let millions of people bang away and refine your algorithms over time based on behavior. It's not deductive science. It's more sort of inductive.

I'm sure you don't need a primer on how search works. I'm sure most of you use search every day in some form or another, but it bears talking about just a couple of components.

Standard search. Put a search term into Google and that search term is run up against literally billions of cached pages of the Web. In other words, we've already captured pages on the Web before you've even searched that term, because the only way to deliver search results in .005 seconds worldwide is to already have a picture of the Web, the entire Web, on our firewall.

So it raises some interesting questions if you look at the history of copyright, because most websites are copyrighted. There's a copyright statement somewhere. Google's crawlers, the software that goes out and does this work, it doesn't actually knock on the door to those websites and say, Mr. Website Owner, would you mind if we made a copy and took it behind our firewall for the sole purpose of actually being able to point people to your website? Now, website owners can and sometimes do opt out, but the holders to our copyright is an opt-in, so this was kind of a revolutionary moment in copyright law because we were copying unless someone told us not to. And the way to tell us not to, or any search engine not to, is simply to apply some code that's universally recognized, robot.txt, and that turns away the crawler.

But if you thought about what the eventuality would be if we didn't have that ability to crawl without asking for permission, the Web wouldn't exist today. The Web would be – maybe permission-based it could be a million pages out of those many, many, many billions that it is today. So keep that in mind when we talk about Book Search.

Once you're looking through the query, we've also done one other thing on the back end. We've determined relevancy. Our algorithms have determined relevancy. The piece of the algorithm, which is about 2230 components large, but the core that everybody knows about that was revolutionary is called page rank, and as scholars and academics, you implicitly understand the concept behind page rank, because Larry and Sergey, when they were at Stanford, ripped it off from scholarly journal publishing. They said, how can we really determine what's relevant? In

relevant? In the old days of Alta Vista and Yahoo, all you had to do with a website was pay Yahoo a load of money and you'd come right up on the top of the search results. But people quickly began to distrust those search results. So they said, OK, how do they do it in the journals world?

Somebody cites your paper. Maybe many people cite your paper, and that's going to actually mean that that particular paper or article is fairly relevant. But what if the person citing that paper – one of the people – was a leader in the field, him or herself cited by hundreds of others? You, in essence, inherit some of that value, that relative value for your own paper. That's exactly how page rank works.

When you see search results, one of the key components is who's linking in to that website and what's their relative value in the world? So if the *New York Times* deems it appropriate to link into your blog site or link into a paper that you've written or link into a publisher's product page for a particularly scholarly book, you better believe that that page is going to come up a lot in search because you're inheriting some of the relative value of the *New York Times*.

So once we've determined that, gone through that analytics in that nanosecond or so it took to give you a result, we provide snippets. And what snippets are are simply some – your search term in bold and some anchor text surrounding that search term to give you a better idea – and solely to give you a better idea – of whether you want to click through to that website or not.

We're not there yet. We've got a long way to go, but we're pretty good and we're better than most. So the question is – and the question was back in 1996 when there was no logical or apparent way to monetize this great technology – the question was, how was Google going to make money on it? In the early days, they thought, well, maybe we can sell search technology to corporations. That didn't work. And then they came upon sort of the idea of applying some of the same sorts of technology of relevancy to advertising.

So you're all familiar with this page. You can't see this but it says, Toronto tourism. When you search Toronto tourism, you're going to get on the left side what we call organic search results. So can't game the system for organic search results. That's all based on relevancy. The top results are basically rich in page length and very relevant in terms of the information it provides.

Now, on the other side, sponsored links, you absolutely can buy your way in, and people do. One of the two major programs through which we make money is Ad Words. Our software, our technology, delivers contextual ads against the search term. Yeah?

M: (inaudible) you said can't game the results on the left side. (inaudible) recently where the keynote speech, one after another after another, talked about how there

are lots of big corporations that make their money showing you how to game results and getting results for you.

PALMA: Well, there's a difference between gaming the results and optimizing your website for search. No, it's –

M: Like underwriting and advertising?

PALMA: Could be. But the difference is this. Gaming would be, I've got a back door in. I paid somebody some money and I'm going to appear on page one. Or, in the early days of Yahoo, because the software was so unsophisticated, website owners would put up their websites and in white type on white background, write cars, cars, cars, cars, cars hundreds and hundreds of times, because if the word appears many times on the website, it must be relevant, right? But the website was selling something completely different, gambling, right? That's what I mean by gaming, and that's click fraud and that's something that we're constantly vigilant about fighting in order to keep this thing sound and keep people trusting it.

M: Again, if I could push it a little further. Now, you've given me (inaudible) SAT coaching. You could do better with them. You do better to pay those companies to maximize your (inaudible).

PALMA: Absolutely. Part of what we do in Google Book Search is if you're a publisher on Google Book Search, we'll help you, coach you, not with any special coaching that you couldn't get from a search optimization firm, but just from us hanging around search, we can say, hey, this is what you need to do with your website. We'll just make sure these publishers, who really aren't very savvy with regard to the Web, are doing the right things.

So when we deliver contextual ads, the advertisers area actually paying when somebody clicks on that ad, so they're buying keywords up in an auction environment and what that means is – and I always use this same analogy because it harkens back to my publishing days. Let's say that you published the key monograph on hermeneutics. If you were to buy the keyword hermeneutics, you might be paying – if somebody clicks on that ad, maybe for a book, you might be paying a nickel a click-through. It would be a very effective way to advertise that book. But let's say that you decide you're going to start a new corporation and you're going to buy management consulting. Now you're going up against Accenture, IBM, many other top consulting firms in the world. You might be paying $20, $30, $50 a click-through rate.

It's an auction environment, and the auction basically regulates where you appear on that list. It's not to say that you couldn't get into management consulting for a reasonable fee. It just means you'd be on page six of the search results rather than page one. It's a self-adjusting marketplace.

And believe it or not, that's how we generate the majority of our revenue, one click at a time. So you obviously have to have software to be able to really scale that beyond belief.

So one of the other ways we do this is through an extension of that called Ad Sense. The only difference between Ad Sense and Ad Words is that Ad Words appear on the Google site and only on the Google site, and it's based on your keyword. Ad Sense is a syndication of that program with rich content partners, publishers, if you will. This is actually the key program because we actually do sell ads against book content in Google Book Search in precisely this way.

So now we've got to get a little bit more sophisticated. We actually have to understand – the software has to understand the context of the page now and deliver an ad based on the context of the page. So this is a Look Smart article about diabetes, and on the right-hand side, we've delivered ads for insulin products and other medical products. Now you're getting to a whole other level of understanding and intuiting what people are looking for and what they care about. Yeah?

M: Closer and closer to artificial intelligence? Is that – read between the lines (inaudible)?

PALMA: Yeah, in the sense that through massive scale, millions – last month I think we had 560 million unique users worldwide, and it's literally hundreds of millions of clicks, of keywords entered and queries a day.

So basically, a real rudimentary way to think about it is, I've entered Toronto tourism, but me and 3000 other people a day do that, but most people click on number three and don't click on number one. What's the delta between number one relevancy and number three relevancy and how do we measure it and how do we improve the ranking based on that? That's the artificial intelligence. That's kind of where the learning occurs.

M: By the way, that does go back to your inductive versus deductive a few minutes ago.

PALMA: Correct, correct, exactly.

So those are the differences. And this represents a vast majority of Google's revenues is based on advertising. We don't sell content. So if you wonder what industry we're in, this is it. It's about $295 billion, and a massive amount of it today is migrating to the Web. Search-based advertising – Web-based advertising, but particularly search-based advertising is the fastest-growing segment of the advertising market today. For one good reason. It's extremely efficient.

Who was it? Wannamaker, who said – John Wannamaker said – he was a department store mogul in Chicago at the turn of the century, and he said, I know that half of all my advertising is absolutely worthless, but I just don't know which half. Advertising is the last bastion of unaccountability in corporate America. Well, now it's accountable. You could actually run campaigns and run metrics against those campaigns and know exactly how effective they are and adjust them on the fly. That's the industry we're in, and we're only 3% of it today, but it's changing pretty fast.

One of the reasons for that growth and adoption is that the Web has exploded. These little red circles represent the number of years it took each medium to reach – I'm sorry? can you see it?

F: Yeah, I can.

PALMA: OK. I didn't know if people could read that. Basically, the Internet took one-fifth the time to reach 50 million users as television. And the other interesting thing that you'll note is that with the rise of widespread adoption in the early '90s of the Internet, you saw a flattening of these other lines, but they don't go south. They continue to go north. And the reason for that, there's a tremendous amount of give-and-take between various mediums, television and the Web, the Web and television, radio and the Web. So the Internet is actually driving, continues to drive, some of these other mediums today.

So by now, I've bludgeoned you with this concept that we don't sell content. That's not our role in the world. We play this role of a switchboard between users who are seeking rich, relevant content, and we monetize that activity of pointing them to that content through ads.

Any questions at this point? You had a couple. Understand the business we're in? I should probably get a watch out here so I know what time it is.

Let's talk about the good stuff, the book business.

Some of these slides – I'll apologize – are speaking to the broadly-defined book business, and that includes trade as well, but some of these trends are pretty universal.

When I left publishing in 2000, traditional retail was a good healthy 40% to 45% of sales, at least on the trade side. Schools and libraries were more like 30% or 40%, and I'm talking across the board, not just university press publishing. Direct to the consumer, oddly enough, still 2%. Non-bookstore retail was not even on the map, and then suddenly Sam's and the price clubs came onboard, and at least for trade publishers, they represent a huge chunk of the market today. Book clubs were huge. Book clubs were a good healthy 10%, 15% of the market, and they're all but disappearing. And online was maybe 2%. It was Amazon just starting to kind of

take off and BarnesandNoble.com, and then it trailed off pretty quick from there. And that's the fastest-growing – not surprisingly – segment of the book market across the board in all but the textbook area. And even in the textbook area, some of the online sellers have begun to makes some gains.

But the interesting thing about the 13% number is actually artificially low because a vast majority of people – 70% or so – 72% of purchasers actually research all kinds of products online and then buy off-line. So the echo effect of the Web makes that number more like 15%, 20%.

F: That data is also from 2004.

PALMA: Yeah, it's 2004-2005, so it's probably higher today. We just haven't found a way to refresh it yet because there hasn't been another study. So, yeah. You'd expect it to be direct, 15% and maybe even higher. Good point. Very good point.

This is kind of a fun slide direct from Google Book Search. So *Freakonomics* is a very successful books on all counts. It's taught in courses. It's a perennial bestseller in bookstores, good trade book, it exports well. Across the board, very successful book. Not that Steve Levitt was a household name before this book, but he is now. But I just wanted to show kind of the impact of things happening off-line with online sales of books.

It was puttering along. Those bars represent searches on Steve Levitt or *Freakonomics* in Google. So it's puttering along and nobody kind of knows who Steve Levitt is, and then all of a sudden it's released so the publisher does some promotion on it and it gets bumped a little bit. But then he appears on the *Daily Show* with John Stewart, and it spikes. And then it slips down again and continues to spike again after he's on the *Today Show*, and you go up and down depending on his appearances all the way through to the BBC Review, which obviously is going to target a worldwide audience, and now it skyrockets again.

It underscores the fact that there's not any unilateral strategy that publishers, academic or otherwise, should be following. They shouldn't migrate to the Web. They shouldn't abandon their off-line. They need to be on the Web, but they also need to promote off-line in order to enjoy these kinds of synergies.

These, again, are more important numbers probably for trade houses, but it's an interesting trend nonetheless that people now spent as much time overall, in the United States anyway, on the Web as they do watching television. I don't think people are watching any fewer hours of television. What they're doing now is actually reading fewer newspapers and books, etc. Although that's somewhat misleading because we know that they're reading journals and newspapers online as well.

And the other thing that they're doing – and this is really interesting. More and more, if you're in the 18 to 25 demographic, over a third of that online audience said that they usually or sometimes search the Web while they're watching television, so there's an awful lot of multitasking going on here as well, so it's very difficult to kind of figure out what's happening.

This slide, in my pitch to publishers, I put up there for one reason and one reason only. Publishers' marketing spend, their advertising campaigns, their publicity campaigns are left far behind this trend. They are still primarily advertising and reaching out to book reviewers in journals and magazines even though their potential users and buyers are going online.

This is another slide that harkens back to my publishing days. For 17 years in university press publishing, all we did was wring our hands over the fact that there were too many books being published in the world, and how were we going to reach the market, and what do we do about it because we're losing market share. Well, we better grow the list. We better publish more books. Every publisher was going through that same thing in their minds, and therefore – this is also 2005 numbers, 175,000 new books being published in the United States alone. There are now, according to Bowker, over 9 million books in print worldwide, and you can imagine how many times that there are not in copyright or not in print. So how do you reach this marketplace today with so much noise in the marketplace?

Well, the good news is that Google actually refers over 60% of traffic to book-buying sites. And when I say traffic, I mean traffic to Amazon, Barnes & Noble, author sites, author blogs, publisher sites, all of it, book sites. And what that means is, potentially here, we're growing the pie. We're taking someone who's seeking rich information and very specific information on the Web and we're converting them into book buyers. Some of that traffic is an ad placed by Amazon. Some of it could be even a product page from an Amazon or the publisher. But lots of it originated from someone who was searching some very, very unique term and found a book and discovered a book. That's kind of what Book Search is all about.

I studied economics when I was an undergrad, and everybody knows about the Prado Principle. You sell 80% of your product to 20% of your buyers. And pretty much every market in the world follows this as a law.

The Internet came along and changed the rules of the game, because what the Internet did was allow for providers of books and magazines and rich content of all types to be able to extend demand down the tail. What I mean by that is, Harry Potter – if you look at Harry Potter as a search term, it's going to be top five every day. Millions and millions of people searching Harry Potter and Britney Spears. And then you get down to Peruvian orchids and you're talking about very, very specific and unique terms. I think it's up to 35% of the queries that we see every day of the millions of queries are unique, meaning one person in the world queried that term. And what's really unique about it is we're selling advertising against

those terms. So what does that mean? That means we're monetizing those very obscure terms down the tail.

And for publishers, particularly scholarly publishers, we're extending that demand towards books that are older, books that are deep backlist, books that are obscure, books that are monographs, because the power of search actually to discover those books. In the past, you'd have to build a bookstore as big as Amazon, which is over a million volumes. A typical Barnes & Noble carries 100,000 books. So Amazon's over a million books. But even if you could build that store, how the heck are you going to find anything? Everything's spined out. There's no search.

So you have the dual processes here. The incremental cost of adding another product to Amazon's Web pages is effectively zero, so you have no stock issues to speak of. And you now have search, and in the case of Amazon, you have social interactive forces at play to drive more demand. Those two things are revolutionary and they're unique to the Web, and they've changed the rules of the game.

F: Did you actually say that 35% of all search terms are unique terms? Is that what you just said?

PALMA: Are unique terms.

F: In Google.

PALMA: Meaning one or two people worldwide entered those phrases.

F: Is it in general Google search terms or –

PALMA: Yes. Yes.

F: That's amazing.

PALMA: Not month in and month out, but we've had months where that's been the case. When you consider the entire globe and what they're searching on, you've got some pretty micro little languages, and you've got unique –

F: You're saying on average?

PALMA: On average. So that's the potential here.

M: (inaudible) one time (inaudible)?

PALMA: Yeah.

M: Explain the tail real quick. What's the tail representing? The very low level of –

F: Is that the long tail theory?

PALMA: Yes. This is the long tail. There's a book by Chris Anderson of the same name, called *The Long Tail*, and all he did was take the Prado Principle and put a nice little curve against it and speak to it in economic terms, and he's got a bestseller on his hands.

The long end of the tail are simply represent selling just a few copies of many things, so you're increasing unit sales in a small way but across many products. And the Web and this effect allows you to extend that tail almost infinitely. So publishers in our program are already seeing that. Books are not going out of print. They're taking books that were out of stock indefinitely or some publisher experimenting with putting books back into print, and actually driving enough revenue off of many of them to put some real, serious margin back in their pocket.

And by the way, typically, books and other products that are in the long end of the tail are books that are older and products that are older that have earned out, and the margins on those products tend to be much higher. So we're actually not only driving demands for these content providers, but driving demand across products that have higher margins. It's like an author royalty situation. If you could sell more of books that are five, ten years old that are in their second, third printing, your royalty rates would typically go up. So anything that drives demand down the tail for backlists is in essence driving demand towards a higher royalty rate for authors. Not across the board, of course, but that could be the case.

Just a little brief foray into e-books. There's a lot of talk about e-books. I was just at a conference of booksellers in Australia and the keynote speaker got up and talked about the explosion of digital warehouses in Europe, how we've got five, six, seven players – most of which are owned by Amazon – coming onto the scene and developing e-book platforms to help booksellers sell books. So what's the market cap right now for e-books in Europe? He said, oh, it's really small. It's really, really small. E-books are not on the horizon. We don't believe that e-books are on the horizon. Too many people are taking the analogy of music and what's happened with Apple and iPods and tried to shift that analogy over to the written word. But in fact, books have all kinds of different use cases.

You listen to music through an iPod, it's basically the same. It's classical music, it's rap, it's jazz. It's the same experience. Plug it in an listen narratively through a song. Books are completely different, depending on the book. If it's a reference book, you might dip in and dip out. Maybe online's good for that, search. A piece of fiction, you read pretty much all the way through. And textbooks and professional books is some combination of the two. You're dipping in and out. There's no device that's going to – in the near future – ride into the scene like the man on the white horse and capture the entire book market. It's just not going to happen, at least in Google's view.

What's the alternative? The alternative is to leverage technology where it should be leveraged, which is on the discoverability side. Leverage the technology of the Web to drive more print sales.

M: (inaudible)

PALMA: Yeah?

M: (inaudible) students, second year of college, could indeed walk around with a little BlackBerry and have those page numbered 81 to 121 out of the textbook right there in the BlackBerry. Do you think that you say about e-books negatively or will you still say –

PALMA: I would say that that's the scenario that has the best shot in the short run, without a doubt. Both digital course packs – and probably that's the number one, if somebody could come up with a model that works. I worked with this company, ebrary, and we were the sort of back end for McGraw-Hill Primis, so we kind of saw that grow. It didn't grow very fast, but it was definitely growing year over year. But that's the best-case scenario, without a doubt.

Vital Source is another company now. They tend to focus in on the sciences. Boston University dental students, they have their entire – what is it, three years – of textbooks on a computer, basically, to this Vital Source interface. That kind of makes sense.

M: The other type of model that I've heard that seems to have potential is not so much, oh, when the technology changes, the books become available, but the book (inaudible). It's almost as though, OK, what Google is doing, what iTunes are doing, don't change that. Let's change our books to fit that model, and I've heard that argument as well. So forget textbooks. We're only going to publish chapters. And then you're going to grab a chapter from this book and a chapter from that book, and maybe chapters are too big. So in the way that you listen to a single podcast for less than five minutes, you might – and it's almost like turning every book into the traditional reference in doing that. While that might not work for fiction (inaudible) work for all the models. I can't see searching for definition in this way and (inaudible) a daily definition of a word. I've heard that argument as well. But of course, there's a lot of reluctance in the industry to start doing that with a traditional book.

PALMA: Actually, the fear is the same fear that is kind of realizing itself in the music industry.

M: (inaudible) part of an album.

PALMA: Yeah. You used to be able to sell the album for the one hit song. People would buy up the album because they wanted to hear that one song. Same thing. You start breaking up chapters. Now you're selling for $3 what you used to have to drive $60, $70, $100 sale. That's their fear.

But if they could, in fact, garner back some of the used book market revenue, which they and authors see zero of, it could balance itself out because you've created sort of the ease of use. You guys could start putting links into your Web CT accounts and they could actually be monetized through some kind of interface so that users could actually have a pretty seamless way to go right through the syllabus in precisely that fashion. So I think that has some merit, without a doubt.

Book Search today. We tipped over, recently, a million books in the program, from the partner program, from over 10,000 publishers worldwide, and that 10,000 is definitely a long-tail number, because some of those publishers are pretty small, and it's worldwide, as well.

Much of the growth that we're actually seeing today is overseas and South America. We have people that do what I do in London, in Russian, in Japan, in China, and those parts of adoption on the publishers side are kind of growing pretty fast right now. We're now live in over 70 countries, and we're localized in about nine, meaning that we've actually localized the Book Search interface in many of those languages.

As far as the books themselves in the program, we're in pretty much every language under the sun. There are a couple that stump us in terms of the digitization process and the OCR process. We haven't figured out how to do Arabic yet. We haven't figured out how to do Thai. But they're really a handful. Our guys have been pretty good about knocking those over when we come into challenges.

So how does it work? Do you remember how we talked about Web Search working? It's exactly the same. Consider every word on every page of a book, those million books, searchable. So the page of the book is a Web page and your searching, your search query, goes against all of those pages. You're given search results. The only difference between Web and Book Search results are that you see a little book jacket there. Otherwise, exactly the same. You've got some anchor text surrounding your search term, which obviously is going to change.

A publisher asked me, well, who puts those summaries in there below the search phrases? And I said, the user does. He said, what do you mean, the user does? I said, because those summaries are really the anchor text surrounding the search query. Because the beauty of this is, what the user could be looking for, the information that they seek is probably not in the metadata, and up until now, we've had to search author, title, description, and that's all we had. Now, what they – we

may surface a book because of what's on page 175. May sell a book because that's the only book that covered that topic. Yeah?

F: Is this a different search category, (inaudible) search (inaudible)?

PALMA: Correct.

F: So on that menu at the top, there's a Book Search?

PALMA: There is. And we've just recently done something that's the start of something pretty revolutionary, that I'll talk about in a second, regarding all those indices. Google has a bunch of indices. You've got maps. You go to Map Search. You've got Scholar. You want scholarly journals, you go to Scholar. You got books, we've got Books. And you search against those. What we intend to do is blend all the indices so that we make two decisions – is this the right information, and is this the right information type – and deliver results. So books are going to come up right alongside websites, and that's already starting to happen now. When that happens, I'll put a book up against a website any day of the week. That's where books are finally going to compete on an even playing field with websites.

F: I just (inaudible). At the top where it says word and image and so forth, what does it say for Book Search?

PALMA: It says Books.

F: It says Books.

PALMA: Yeah. I'll show you that in a second.

F: (inaudible).

PALMA: Up until now, you would have to go to the More tab. Actually, books, no. Books had some real estate at the top, I think. But in some countries, they don't. In some countries, Books doesn't get the real estate at the top, which is very dear to Google. You don't get up there very easily.

But now you can actually filter results, so if you were just searching Google Book Search, which you can do, books.google.com, and search across all these books, you're going to get search results that look just like this with sponsored links on the right-hand side. And then you click on one of those links and you're going to actually be brought to the first page in the book that contains your search query. It could be page one, but it could be page 192. Wherever that search query first appears in that book is the page you'll be brought to, and only 20% of that book will be browsable at that point.

The key term here is that it's not any 20%. The user doesn't get to go in and say, ah, well, I know I've got 20%. I think I'll go into chapter 12 because that's all I need for the course, and I got it, because it's 20%. No. A very small portion of the book is actually delivered programmatically. It's sort of a sample block that we've developed programmatically, and it's just a few contiguous pages of the introduction and then we roll them through to chapter two, and a few fewer pages as we roll through chapters, and then we show them the index, and then we show them the back cover and they're done.

The only way for them to actually see more pages at that point would be to enter another search query. So somebody said, well, they could come back to the book and see more pages, right, by just entering more queries? They absolutely could. If they had the book open in front of them, they could actually pick very unique terms on every page and search against those pages in order to get contiguous pages. They'd still be blocked out after 20% of the book, and there is part of every book that we dynamically block, blacklisted pages, just to kind of underscore to the user that this is not a book-reading program.

F: Even if the search terms are on them?

PALMA: Even if the search terms are on there. So the –

END OF SIDE

M: What was that term you used? The blacklist?

PALMA: Blacklist, yes. We've got whitelisted pages that are always there. It's kind of a sample block that we create. And then you have blacklisted pages which are kind of a Swiss cheese effect.

F: What's the percent? I use it a lot, Book Search a lot.

PALMA: Do you ever hit them?

F: I've never hit a blacklisted page.

PALMA: You probably just don't hit the barrier. But if you use it enough, you'll (inaudible).

F: What percentage is it?

PALMA: I can't tell you what the percentage is.

F: Oh, you can't?

PALMA: No, but it's a portion of every book. And then it's very low resolution images. These are for those – go ahead.

M: Are the books on here only books that might (inaudible)?

PALMA: That's a good question and we're going to get to that, too. So just to kind of complete the thought here. You can't print the page, you can't copy-paste, you can't do anything with this content other than read it online, and then we provide buy-the-book links for the publishers.

We've also started to build out these About the Book pages. Basically, they're product pages, some of which we take third-party – it's a bad example for this crowd, heavy metal. But basically we've built out these pages to kind of give people a better sense of how the book fits into the world, so we're now starting to include reviews that we spider from the Web and we pull them in. We have scholarly articles that reference this book or cite this book will be listed, so you can bounce back and forth between Google Scholar and Google Book. So there's some pretty neat stuff happening on those pages. And we have direct correlation data that shows the longer people spend on these pages, the absolute higher proportion of them click on a buy-the-book link.

And by the way, for those publishers who are experimenting with opening the entire book up, that number just keeps going up. The more pages people view, the more, a higher proportion of them click on the buy-the-book link. It doesn't go down.

F: Is buy-the-book taking them to the publisher's direct sales site or is it taking them to Amazon and Barnes & Noble or –

PALMA: Both. The publisher is always the first link, and it goes to the actual book. So if they sell, if they do e-commerce and they want to do e-commerce, and many of them don't because they can't fulfill a single book in 48 hours or at least fulfill the expectations, but if they do want to, they're always the first link. And then the other links are – well, depending on the territory. We change them. It's not always the same. So it Australia, it's not going to be the same as the U.S.

Our whole intent there – we don't get any referral fees whatsoever for providing those links, so when somebody clicks through to buy a book, Google gets nothing. And we do that for a reason, because we want to show people that we're actually about sending the user to a place where they can actually get hold of this book.

If we could get stock information and dynamically generate those lists, we would, but booksellers won't give it to us. So we just take the best guess we can based on our market knowledge and provide a listing of those retailers where we feel the user's going to have a good experience trying to locate that book.

M: But that's got to be temporary. You're going to want make money on (inaudible).

PALMA: No, it kind of fits in that same category of organic search results and buying your way to the top. Once we start doing that, a) we're going to get a flood of booksellers that scream bloody murder in the industry rags that why aren't the included, why can't they be included, why are you giving preferential treatment to other retailers. We just don't want people to think that somebody else is gaming that list.

We're also starting to do some neat things with our other products, what engineers call mash-ups. In the UK anyway, we're experimenting with the ability of the user to type in their ZIP code and get a local bookseller, generated through Google Local and Google Maps. Ultimately, again, we want stock feeds. We want to point them to exactly the bookstore in their neighborhood that has that book, and that's going to take a little time to get there.

We're also – I'm not sure what the kind of value is here, but it's pretty cool. On those product pages, if a book is rich in place names, you'll see a map on those product pages as well that maps every place name mentioned in the book as well.

F: (inaudible) now indexed? (inaudible)?

PALMA: You mean the percentage of the books in print that we've indexed?

F: Yes.

PALMA: Well, it's a million books is the public number, and there are about 9 million books in print worldwide.

F: And whatever happened to all of the lawsuits and everything? Are they still in progress?

PALMA: Yep, and we'll talk about that. That's the library program, right?

So the big news is there's two ways to get to books now, Google Web Search and Google Book Search, and we talked a little bit about that. So Universal Search is Google's attempt to merge their indices, and when that happens, if we do it right, it's going to be a major bonanza for book publishers for sure, because now books will come up right alongside websites.

And today, they tend to come up most often, almost guaranteed – if you typed an actual author-title or a title, it's going to be the first result now, often with a book jacket. Because we're feeling pretty good about relevancy when you type it in.

M: So if we're not aware whether our earlier editions are in your database, if we were just to go upstairs to our hotel room and type in a few words that we remember

were unique on page 172 or something like that (inaudible) probably didn't do, we'd find out (inaudible)?

PALMA: Well, you could type in the author-title. That would be the easiest way to get to that book.

And there's a lot of dialogue about whether we include previous editions, whether we don't include previous editions, and there's schools of thought in various disciplines. If you folks have opinions, I'm all ears.

So *Fast Food Nation*, right? If you were to type in Fast Food Nation, you'd get a book jacket. But if you just typed in popular dogs in Japan, one of your Web results is going to be a book called *Doglopedia*, *Planet Dog*, and it's going to appear right alongside websites because popular dogs in Japan are in that book, but sort of in the middle of that book.

Our user studies actually found that when we put a jacket up against the websites, people tended not to click on it. People are just so comfortable with Google that you throw anything, any curve ball at them, I don't know what that is. It could be an ad. Better not click on it.

M: I thought the reason would have been that if you're already very familiar with what the book part is, and if I'm looking for an answer to a query, the website might give it to me quickly. The book site might tell me I've got to buy the book to get an answer.

PALMA: Yeah. That could be part of it, but at least the feeling on the part of the engineers who have done these tests feel it's just literally the curve ball. It's their so comfortable with the interface that if you throw something new, it kind of – unless it's actually querying the title, then they would expect to get a book.

I don't have a lot of real data from publishers to share because they don't share it with me. The one thing that we can show publishers is that a buy-the-book click happened. We don't have any data on conversion. Did they sell a book based on somebody clicking through? We know anecdotally from publishers that share that in fact there's a pretty high rate of buy-the-book clicks and there's a pretty high rate of conversion. So I can spout this stuff because we just had the Book Expo and we had a panel and these guys said this publicly on a panel, so I can say it here.

Springer-Verlag, obviously. I've got two professional publishers, but that's who were on the panel. 99% of their books were viewed at least once every month, and they've got 20,000 books in the program. 75% had buy-the-book click, and 25% of the buy-the-book clicks are – and this is important – for books that older than ten years old. This is in science, medical-technical, and these are books that over ten years old, and they're now seeing long tail impact of Google Book Search as a result.

Same thing on the McGraw-Hill Professional side. 91% of the books were viewed. There are about 4000 books from McGraw-Hill Professional.

60 million book pages have been viewed in the past year, and 22 of the top 30 titles viewed were older than 2004.

What's interesting about STM publishers is there's a much higher proportion of people that click on the publisher link to buy the book, because there's much more identification with the publisher. Whereas, if you're in the trade world or textbook world, the book itself or the author is the brand. There's no brand, so they tend to click more on Amazon or –

F: How does the ISBN fit into all this?

PALMA: It doesn't really fit. If you were a librarian or you were a bookseller and you were actually looking to get a hold or information about it, you could go to an advanced search in Google Book Search and punch in the ISBN, but most people don't. Unless you were a professional in the book trade, you probably wouldn't punch in an ISBN. What did you have in mind?

F: Well, usually and traditionally, when I want to buy a book that's not in the bookstore, I just go to Barnes & Noble, give them the ISBN. They order it and they call me up when it's there. If I go and pick the book up, (inaudible).

PALMA: Right, yeah. You could go that route. It's a kind of circuitous route because you already know what you want, so the intent is not necessarily to point you to things that you already know you want. It's kind of to point you to things you didn't know exist.

F: That was always the positive identifying (inaudible) of the book, the ISBN (inaudible) right book.

PALMA: Oh, the ISBN is definitely there in the metadata, without a doubt. And in fact, when they click through to a retail site, they're brought to that very book based on the ISBN. They're not brought to the home page of the bookseller or of Amazon. They're brought to that very book.

F: So there's no risk (inaudible).

PALMA: No.

F: (inaudible).

PALMA: Right. I think actually Amazon does list other editions, but it would have to happen on their side. It doesn't happen on our side necessarily.

So this is the library project. Three minutes. OK. We talked about the controversy. This is what the controversy's about. We have 16 library partners worldwide, everybody from the University of Michigan to Harvard to Oxford, New York Public and increasingly, some foreign libraries in Spain, and I think we had our first in Belgium recently.

The intent of that program is to put books in the program that by and large have some unclear status. The vast majority of books that we scan from those libraries, the copyright status is unknown. They're orphan works. However, if they are in copyright, we won't show those pages, so those excerpt pages that you see in the partner view, you won't see on a book that's in copyright that we scan from a library that's not in the partner program. There, you would see a snippet view. You would see – literally, it looks like a couple of torn lines from the book, two or three actual sentences from the book surrounding your query, and that's it. Obviously well under the bar for fair use, even though those books are not in circulation.

So we're providing access to books and hopefully driving some traffic and discoverability towards books that haven't seen the light of day. That's kind of the intent of it. If your mission statement is to organize the world's information and make it universally accessible and useful, we've go to find a way to get these orphan books into the index.

But that's the controversy, right? Because we're not saying, hey, we discovered some books, Mr. Publisher, that you let go out of print. We want to scan them only for the sole purpose of pointing the book so people can get them. We don't show that book. So that's kind of the rub at this point.

The various views you can get is you can get a full view on a public domain book. If it's in the public domain, we show the whole book. If it's in copyright but out of print or not in the partner program, we show this little snippet view. And for the books that are in the partner program, we show the 20% limited preview because we have the agreement with the publisher to actually show those books.

Then there are some books that we decide, even though we can show more, we don't because they're – maybe they're rich encyclopedic type of content and we only want to show a small portion of it.

F: What's the snippet view?

PALMA: The snippet view is the book that's in copyright but out of print that we scan from a library. I shouldn't even talk about it being out of print. It could be in print. But we know that we can't show pages because we don't have the rights to show pages, so we only show your search query with a –

F: Is it (inaudible)?

PALMA: It's out of print but it's in copyright, so it might have been published after 1923, or maybe it was published before 1923 but we found some photographs in it and we don't know what their status is, so we just err on the side of caution.

F: (inaudible) as partners (inaudible).

PALMA: That's right. Exactly.

F: And no preview is what?

PALMA: No preview is really books that we have decided we don't want to show anything of because maybe it's a richly encyclopedic in some way. So we want to show metadata but we don't want to actually show the book. The book will still be searchable. It'll still come up in search results based on a query. You just won't be able to see words.

Then, if the copyright holder – even though a book may be out of print, they could send us a list at any time, even if they're not in the partner program, and we will block those books from entering into the library program entirely.

F: If you could go back one. No, back the other way one more. In full view, what are the restrictions? The book is all there, it's all scanned. It cannot be copied? Can't be (inaudible)?

PALMA: No, public domain books can actually be downloaded. Not all of them, but most of them are downloadable and fully viewable throughout. You can't print them, but –

F: Let me just see if I understand. I'll use an example. I found a book online because it's out of print and the author is dead and (inaudible), doesn't seem to have any heirs, and it's (inaudible). They probably (inaudible) Google. If it were full view, I could pick a chapter and save it or copy it?

PALMA: You'd be able to save the whole book as a PDF, for many of those books. But that's with the proviso that it's actually out of copyright.

F: Well, I would know that. Google would know that.

PALMA: Right, exactly.

F: Now this leads into my question. How do I know if what I'm looking at is a full view or if it's a preview?

PALMA: Because it messages that in the search results. Right below on the search results page – and if you do a couple of queries, you'll see a – right here, you see limited preview up there. You would see full view. Here, you see snippet view. It actually shows a little piece of torn paper, so it kind of does message you.

Admittedly, it's strange having to use various views, but people are starting to get used to it and we're now providing the ability to filter by view, so you could look at only the books that are full view, only the books that are partial view, etc.

M: Say that I'm an indexer so I'm working in a closed environment, one book, and it's all human interpretation of relevance versus sort of the popularity, very scalable style that Google uses. And I know that that's one of the discrepancies when you use the term relevance, and I've heard a lot of research librarians who will say, it's not relevant. It's simply popular. And at certain times of (inaudible) –

PALMA: Those don't mean the same thing, though.

M: They don't mean the same thing, but in certain environments, boy, they come really close, and in other environments, they may not come as close. I'm curious in the case of a book. You're searching the book for content, and I know as an indexer, my interpretation of every word is do I want to index it. I would consider, assuming I'm at least a halfway decent indexer, it's probably of greater relevance than just a full text search. I'm wondering if you ever leverage the text in the index as sort of a boost toward – a human being thought this was valuable, and treat it like a meta keyword, in a way, to boost certain things.

So if you've got an index entry with lots of page numbers and lots of subentries, when somebody searches for that word, that book should be more relevant because clearly, the indexer thought it was relevant. I'm wondering if you can leverage the strength of something like that.

PALMA: I can say with all confidence that I don't know. Because there are hundreds of different factors in the relevancy rankings and I just don't know. It would seem logical that it would, but I don't know. We do extract key words, as well. It would interesting for you to look at, to look, to –

M: Well, all the examples, when I perform a search, I'm not seeing index pages pop up and I'm curious (inaudible) cut out of (inaudible).

PALMA: No, no. It's in there. It's in every one of those books.

M: Can I search the index or is the index just a feature that's provided when they match?

PALMA: It's just a feature.

M: So I'm matching text in the book and then I can look at the index for that book if I have access to it.

PALMA: Right. Exactly. And you could go to the product pages, the About the Book pages, and we've extracted key terms. So it would be interesting from an indexer's point of view whether we were accurate. Like, if you were to peruse the book, say, hey, that makes sense the way they –

M: I would expect some accuracy (inaudible), even if it's just a hit and miss, you're bound to (inaudible) quite a bit.

M: Believe it or not, I think we found one interpretation, the end of the Internet (inaudible). So I want to thank Chris for – and by the way –

F: May ask one (inaudible) question?

PALMA: Sure.

F: Because it's about – it's something we haven't talked about. I just now looked up my book and the second edition came up but the current edition is the fourth edition, and I was just wondering how does Google resolve multiple editions?

PALMA: They would all come up next to each other.

F: They did not. The second edition came up, came up first, and then references to some other editions, but the second edition came up (inaudible).

PALMA: Are we certain that the most recent edition has been submitted by the publisher?

F: I don't know. Is that (inaudible)?

PALMA: Yeah. That would be my question. Did the publisher hold back the most recent edition?

M: I want to thank Chris.

(applause)

PALMA: My pleasure. I'll be hanging around if anybody has any other questions.

END OF TAPE