# Integrating Discovery and Access for Scholarly Articles: Successes and Failures

Anurag Acharya
*Google Scholar*, darcy@google.com



Anurag Acharya, "Integrating Discovery and Access for Scholarly Articles: Successes and Failures" (2012). *Proceedings of the Charleston Library Conference.*
http://dx.doi.org/10.5703/1288284315073

# Integrating Discovery and Access for Scholarly Articles: Successes and Failures

*Anurag Acharya, Founder and Lead Engineer, Google Scholar*

*The following is a transcription of a live presentation at the 2012 Charleston Conference on November 8, 2012. Video of the session is available on the Charleston Conference website at http://katina.info/conference/video_2012_acharya.php.*

A brief note as to where I come from: I am a lapsed academic. I grew up on campus. I stayed many, many years on campus and had no intention, in spite of my advisor's best efforts, to leave. I grew up in an academic household; your view is very academic. Turns out, one of my colleagues was VP of Engineering at Google and said, "Come spend a year with us, learn what the new world is like, what the services world is like, and then go back and do your magic." Seems like a fantastic offer, as it was indeed a new world. I went there, and at the end of the year I was running Google Crawl, and I could not convince myself that anything I could do, if I went back, no matter how smart I think I am, would do more than making it possible for everybody to find things. Everybody including the people who I grew up with, a small town in India. When I ended up at Scholar, was after I ran Google Crawl for several years, it's kind of a very intense life at a smaller company, so I took a break, a sabbatical so to speak, to build something that I, in my mindset, would want to use. That has now grown to occupy the last 8 years of my life. So in that sense, I am fortunate, and I will talk about some parts of it.

We started with a simple idea. Everybody must be able to find everything. It's not been the case, and that has been considered normal. Discovery has often been tied to access, because the place you get access is the same place that provides you discovery. You go into an OPAC and you see what the library has for you, and that's considered a good thing. Discovery is often tied with an area, a genre, a particular class of content. Data, journals, publications have been grouped together in discipline-specific databases, which are very useful in some contexts, but limiting in many other contexts. And discovery has long been tied to source. If somebody has blessed you, you can be found, and if you happen to be someplace else where you are not considered to be worthy, things that you said were not worthy to be known.

The problem is answers are not limited by area. There is much interdisciplinary work today and that continues to accelerate. Insights are not limited by a geographical area. Some people aren't the only people who know what the right thing is. An example, the high-yielding cereals that saved a billion people. The Green Revolution came out of work in Mexico and in the Philippines. I grew up in a place with people much smarter than me, but then I have the resources to do many things. They are still struggling. If they had the ability to contribute, what more could we make possible? There are major challenges facing us as a species, us as a planet. Problems that have been with us forever, and the problems that we have actively with much effort created for ourselves.

But if there are major challenges, there are also major opportunities. The connectivity allows more people to contribute, more people to build a shared cathedral of science, but they can only do so if they know what their peers have done. That's what we have started with. That's what we have tried to achieve. One place that you can go to find all scholarly literature: all areas, all languages, and to the extent I can do it all the time. All research from everywhere, no matter where you are. You could be at MIT or you could be at a small college in India. And the other thing we did at that point was to index full articles, not just abstracts. This may seem like not a big deal today, but when we started down this route, where we were basically told, "You are creating so much noise for everybody!" That is not the point. The point is not everything that is important about a particular piece of work is known by the person who did it. So you can describe it succinctly in the titles and the abstracts, but sometimes the magic happens by something you did along the way: a limit that you establish on the way to the theorem; a mechanism of analysis that you did to allow you to do your work. Your work may not be important, but your mechanism ends up being seminal. And it



http://dx.doi.org/10.5703/1288284315073

needs to be free for all users no matter, no barriers, you could always find things.

So where are we? We have built the largest scholarly search. At this point it includes every source that I can reasonably think of, and some sources may be borderline scholarly, but that is the nature of trying to do everything. All languages with significant amount of scholarly content, this includes the East Asian languages, all areas of research. Good or nice. It may not be good. But that's step one. First, and we needed to do this first before we can do anything else, first you need to find and then you need to learn; and titles and abstracts are not sufficient for learning. You need to know the details. My advisor frequently pushed it down our throats. Science is in the details; you need to know what it is about. But actually reading things is far, far more complicated. There are many, many sources and, as you can see, many, many, many pathways of how you could go about, once you've arrived at something that is of interest to you, to actually read it.

So I will talk about some of our efforts, in the context of Google Scholar, of trying to address some parts of the crazy quilt. Talk about access based on subscriptions, some of the efforts that many publishers have done to provide access to free archives. A topic that is very close to my heart is to make it possible for people everywhere to be able to do more to learn, and to be able to contribute to the common knowledge; and what has started out in some fields, and maybe is expanding to other fields, is looking at if early versions of work are a way for you to be able to learn at least something about it. So the overall approach in all of these things is to keep discovery the same. Your result sets are what we can deem to be the most relevant for your query. That doesn't change because first you need to know what is important for your question, and then, depending on who you are, where you are coming from, who you are associated with, what else is known about this particular article, we add additional links to the interface that would allow you to access, or to indicate to you that you have access, or to click through it and get access. To be able to do this, we have worked with a variety of different partners across the whole industry to learn who has access, what they have access to, and integrate that into the search.

So whenever you are doing a search you are actually doing two searches: you are doing a search across the documents to see which documents are the most relevant for you, and then you're doing a search across access information to see what is accessible to you, and then we are doing an intersection for you. So whichever of your search results you have access to, you get additional links indicating that.

So this is what I'm trying to demonstrate is the query “prions.” Prions are the proteins that cause, supposed to cause, the mad cow disease. Assuming you are coming from Harvard. What happens is on the right-hand side you get all of the results that are accessible because of your affiliation with Harvard. You get a link saying “find it at Harvard”. The text of that link is of course chosen by the library. The library tells us what we have access to.

There are several approaches that we have taken down this route, and I will describe them. The first was to basically see what libraries had already been doing in trying to address this issue. This problem is known much earlier than we came by. So libraries have, are people familiar with link resolvers? Let me just give you a brief notion. Link resolvers are sort of an indirection server. It says if you tell it what article you want, it will give you a link to where the library has a copy of it or where the library has licensed a copy from. So we came up with a model with the link resolvers that know what the library has subscribed to. It can export this information in a form that we can periodically pick up and integrate into our index. We worked with all of the link resolver vendors to come up with a suitable format and a mechanism by which they could export this, notify us when a new library joins, or a library leaves, and what have you, so that for a library, all they had to do was a component of a resolver configuration. We launched this fairly soon. As you can see, as soon as we launched Google Scholar, I had a call from some of the libraries as well as from Ex Libris as well, saying, “Hey, hey, hey, did you actually hear about this other problem?” And I said, “No, no,

no, I had no idea." Nevertheless, we were able to move pretty quickly. Over the years, today we are working with every link resolver vendor, every link resolver provider out there, there is a long list.

So where has that led us? It is good to work with, but if you build it do they come? Indeed they do. Today over 4,000 libraries worldwide are setting up link resolver–based integration with Google Scholar. Institutions of all sizes everywhere, I could give you graphs and stuff but it tells you nothing more than this particular piece of information. So why did this work? The reason why I'm saying this is because there will be other things down the line which will not have worked and we'll have to examine each of them in the same way. A big part of this was to make it really easy, once the library decided it was something that was important to them, the steps that needed to be taken were relatively small. But another, and very big issue, is that the libraries that have set up link resolvers are already thinking about this problem. They have Mindshare; they have resources; they have people thinking about this. It becomes very significant when you come down to other approaches that we have gone to, as to what succeeds, what doesn't succeed. But 4,000 libraries? Well, there are so many more libraries, you call that a success?! I'd like to, but there is much more to go. So link resolver adoption is kind of uneven. It is higher in the UK and the US; it is sort of spotty in other places.

So we said okay, if this isn't there, let us try to do something more. Ex Libris was fantastic in setting up a special link resolver which would be hosted free, and easy to configure with a small number of options. It could be more, but it's lovely of them to set this up. And then we went out. We talked to consortia, we talked to, I don't know if people know eIFL? eIFL as an organization funded by Soros' Open Society Institute that tries to help library organizations in Eastern Europe and Africa and in other parts of the world and in Southeast Asia to sort of work together, both to get access, as well as to bring library practices up to the current level. We created step-by-step "how-to's," screenshots, translations into different languages, we did presentations, we did webinars.

What did that get us? Not much. Why? Now this is my understanding of why, and I would always love to hear more from others who may have a different understanding of what it takes to make such a thing possible. Smaller libraries are already overstressed, not just in terms of money, but also in terms of people and what kinds of things they can think about. How many nails they can possibly hammer. There's not enough time for people to explore opportunities, even if they are pushed at them. We had these things mailed, in some countries, we had these mailed to every single library, but no. We considered trying to do this in some sort of half-centralized way in different places. Can a consortium do this? Can a group of libraries do this? Can some other way make this possible? No. Turns out there was just way too much variation, more variation for the smaller libraries than for the larger libraries, and link resolvers aren't set up to do this in an easy way. So we said, "Okay, we have hit our head on this wall for 2 years. Can we do something else?" So there are two sources of this information: libraries, and those who provide the access. Can we knock on that other door and see whether it can be done better? Same mechanism, it controls the same information. Who has access to what? So we have had some success. There's a list of several of our partners that are participating in this, and several other partners are currently in progress.

There are different approaches that our partners have taken, and I want to examine them to see how each of them has done. So the first approach is actually not that different from the link resolver except that it doesn't require a link resolver. It says every library has to explicitly make a request in some form saying, "Please turn this on for me." The advantage, of course, is that you are taking advantage of existing relationships. People are already trying to configure things to some extent and maybe you can leverage that to cover more than you would've been able to cover otherwise. And indeed, to some extent, that is the case, but what about structure? And the problem is that the smaller libraries don't actually have as much time to think about these issues, or to decide which of the many options that are available to them should they be exploring? And they would have to

do this for every one of their providers. So now we have scaled the problem in a different direction. This has made quite a bit of progress, but has limitations similar to the link resolver approach.

The second approach was to allow consortia to opt in, saying consortia could say, “Yes, this is something my members would like. Yes, the content that we are paying so much money for, please make the discovery and access seamless.” It seems like a no-brainer to me, but it was a request that had to be made. The advantage of this is, of course, is that it allows library organizations to help the smaller of their members who may have less resources and be able to do it in a scalable fashion. You can see the numbers up there in terms of how many consortia opted into this, again for some of our partners. The question was, “Why did it work so well internationally?” There are 37 and 5. Well, of course the 37 are smaller 37. They are smaller countries, they’re more homogenous; so there is that one advantage. There is also an advantage of an activist group explicitly trying to coordinate things. eIFL took the lead in trying to coordinate this across all the consortia, to convince them to move, bring them to the table. In Australia, they are just like the United States; every state has a separate consortium. The National Library took the lead in making this happen. They talked to everybody. They told them why it was important. They brought them to the table, and we were off the ground. Why didn't it work in the US? Well, most consortia did not quite see this as their role. This wasn't what they thought themselves doing. Not making requests on their members’ behalf. That is unfortunate. Turns out, some of the consortia were willing to take the leap. There's the five I mentioned; I mentioned some of them up here: the Connecticut, the Georgia, the Virginia consortium, the few others of them were willing to take the leap; and clearly their members benefited. Yes, it is not the role that they’re necessarily seeing themselves in, but if this is a role that they can see themselves in, there is a significant benefit at the other end of it.

The third approach was to say, “Okay, this is clearly a good thing. People who are paying for this would like to make it easier for those they are paying this for to be able to access it.” But of course with everything, some people may think differently. So let's move the default in the other direction. Let's provide this as a service. Let us say that this is now a service that comes with the subscription, and if you don't want this service you can turn it off. By far, by far, order of magnitude the most effective approach. You can see why. And the fear that some people may not want what appears to, at least most of us, to be a benefit turned out to be unfounded.

I will give you a list of at some of the providers that were participating. Why isn't everybody doing this? So we knocked on those doors as well. Suddenly, the response, I got back an entirely desirable response, was that no one’s asking me for this. Libraries are not knocking on my door saying, “Please turn this on. Please make this possible.” You say they might need it, but hey, no, ideally you would want the people that might benefit from it to be able to say this. The problem, however, is the libraries that think about this frequently, and have the resources to deal with it, have taken an approach that already deals with it. They have bought link resolvers. They have worked with link resolvers to turn this on. The others, and a large number of others, haven't, and they are not going to knock on your doors. The natural place in my mind, coming from a different place and not being a part of the library community for such a long time, would have been that the library organization, as if an individual member is not able to have the resources to do it, then a pool would be able to do this. I would still love to be able to explore those opportunities even though we have not always been able to make this happen in the past.

Another initiative, I'm sorry, other than this content that you pay for. Then there are initiatives the publishers have undertaken to provide access to older articles in an attempt to balance, in a judicious fashion, maximizing access, maintaining the continuity and the sustainability of the business model. So articles between 6 months to 4 years old are then, at that point, made available to everybody. Some of our partners have made the p1923, which is copyright-free collections,

available to everybody. The accessible archives, from the point of view of setting this up, are just like subscriptions. They are just subscriptions for everybody. This is what you have access to, and everybody has access to this. So the way to set this up is exactly the same. The mechanism, the implementation is exactly the same. So we came up with a succinct way of being able to specify a moving wall, you wouldn't have to update this on a regular basis. We worked with hosting platforms to make this easy to explore to set it up in a reasonable way, and we have had a fairly large number of partners and journals participating. What you will see, is on the right-hand side of the search results, a link saying you have access to this from JBC. This is the *Journal of Biological Chemistry;* they have a moving wall access.

This integration highlights access that publishers are already providing. There is a huge benefit to researchers and I think it's a wonderful effort, and the publishers who actually do this should be highlighted. Well worth highlighting, both from the point of utility as well as in form of the credit to the publishers. A lot of these journals turn out to be in the biomedical field which is even more important which allows faster turnaround for things that are important for human health. I gave you a nice large number saying many are participating; many still aren't. So I asked myself, "Why not? What allowed one to work and not the other?" It's sort of a circular waiting. Publishers say, "If the hosting platform that we are on supports this, we would love to make this available. They do everything for us." And the platform said, "If we have an explicit formal request from publishers, yes we will put it up." And we wait, and so that's basically what's holding most of the other ones that currently aren't available.

The place where you break this logjam, and Highwire Press was one of the early adapters of this, you see pretty much all publishers who have this particular model join sort of almost in mass. There is a logjam to be broken. A little bit, any place we can break this logjam, I think we can make a lot more of this far more visible in terms of, you know, giving credit where credit is due, and actual impact to researchers being able to access this, find this, follow this up.

So I will switch gears and talk about the developing country access. There are many, many efforts that publishers have undertaken for this. This is in no specific order, and I will describe them, just in terms of describing different ways of doing this. The Highwire Press program which enables access for on a country-by-country basis. All IPs in a country have access. This is actually a very interesting approach of doing this, largely because IPs and institutions are not very well aligned in many of these countries, as well as a lot of the access for most of the students and researchers actually is off campus. You're not provided as much formal on-campus access. And unfortunately, in many places, access to the places where you can "get access" is a source of power. If you know the right people then you can get the key to the right room that happens to allow you to access this. Not that different from a locked library. The very large effort by a very large number of publishers who were now very aware of the Research for Life initiatives that tries to provide a large collection of journals in health, food, agriculture, and other environmental research to developing countries. The JSTOR Africa initiative where institutions in Africa can sign up; the plenary approach requires a password and requires proxy-based access, and I will mention why this is significant.

The JSTOR approach and the INASP approach. INASP is trying to bring together libraries and countries that need access, and publishers were willing to provide it and act as a matchmaker. Integration again is similar to what we do for subscribers. If you are coming from the right country, or you're coming from the right IPs, and your results include results that are being made available by publishers in this fashion, then you will see an additional link that indicates to you that you have access, and that if you click on it there is the pathway for you to get it. I cannot emphasize how highly I think of efforts like this that allow, that ask of the world to contribute to the problem that all of us have.

So what worked? Now, "worked," keep in mind, is in terms of integration of discovery and access. Please keep that in mind. Programs themselves work well in and of themselves. All of them.

We've had this program since 2009. It's a very straightforward thing and fits in very well. It says all IPs from this country, they tell us which countries, and we add the link for those countries; things just work. Beautiful. The JSTOR Africa initiative, we've been working with them since 2010, why did these work? Because the hosting platforms themselves are committed to these, they are running these programs, they're committed to these programs, they want to make this thing work, they want to highlight it, there is a single entity that is willing to step up to the plate and make it happen. And they have shared infrastructure that they are doing for subscriber access, for archive access, so it is less work for them.

What didn't work? And this broke my heart! It really, really did. I knocked my head on every single door that I could think of. The unfortunate fact is that I don't even know why I am unsuccessful here, because it seems so clearly a benefit, and more so because you're coming from a proxy. So much wonderful access is available and we're just kind of halfway there.

The INASP program, in comparison, they can see themselves as a matchmaker role. They did not see themselves in the role of making requests for this sort of information or trying to coordinate libraries. Their point was we make it possible for people to get access, and it's a wonderful thing they do, but they are not willing to take the step beyond that to basically, in effect, be an organization of these libraries. Many articles are available in many versions prior to publication in many fields. Not every field. These are usually deposited in discipline-based repositories around the world. There's different, there's Archive, there's SSRN, there is a whole different bunch of them. What we did was to work with both of the disciplinary repositories, as well as the people who build repository software that libraries use to store the run repositories, to make it easier to index them; and once we index all articles, we group different versions of the same work. It's not necessarily a different version of the same article; it's more for like different presentations of the same underlying research work. What you see in Google Scholar search results is not actually articles. They may sometimes seem that way because there is only one presentation, but many times there's many, many presentations. There's a preprint, there's a conference, there's a journal, there is an anthology, there's all of these different versions all of them get grouped together. You may have access to one of them. We group all of them together, and we link again.

What I was going to show there (referring to a link in the slide) was a query; this is a string theory query where you can see you can get the results, the formally published results, as the normal results plus the preprint versions from archive available in links on the side. They made much progress, but there is so much more that remains in this space to make it possible for people to learn.

What I have listed up there are some of the challenges and some of the hopes. The smaller international libraries are where the challenge is biggest for subscription access. I'm hoping to work with library consortia; I probably will be meeting some of them while I am visiting here. If you are one of the libraries that fall in this category, please also talk to your library consortium to see if they would be interested. It is a small, not very large, amount of effort we can make this possible; we have made this possible elsewhere. For archive, for free archive access, I would like to make it possible for all of it to be highlighted as much as for the ones that we have done. Again, it is not a very large component. We're out of time. For HINARI, and the Research for Life, I would love suggestions. I would like to draw upon the collectivism of people here who have been in this field much longer than I have to see what might work, what might be possible, or what we have done wrong, not to be able to make this possible. I would like to leave you with this: not everything has the impact of Mendelian inheritance, but there is much that does. You don't know which one of them is going to. The more we make possible for everybody around the world to contribute, the more likely it is that we will succeed as a species. Thank you.