Interview of Sebastian Hammer, co-founder of Index Data, by David Weinberger, of the Harvard Library Innovation Lab and a member of the Dev Core for the DPLA on 2/7/2012

DAVID: Hi Sebastian...

SEB: Hi David, how are you?

DAVID: I'm okay, and I have a bunch of questions, almost all really simple. So first of all, can you explain "federated search"?

SEB: Well, the term has meant different things over the years, but I think the consensus today probably means providing a single interface to a number of different information sources by broadcasting essentially a user search and then trying in some way or another to combine results and show them on a single display or screen.

DAVID: So this is very different from, say, how Google works?

SEB: Oh quite, yes. It essentially means going out sort of sending every search out to every one of the different database silos that you're interested in.

DAVID: Why would we do that instead of basically copying into our own site everything that we're trying to index and running the index on it and doing it all centrally?

SEB: So there's a few situations where you can really gain some benefits from that approach and there may be some situations where you don't have another choice. To talk about it sort of specifically, there might be data that you're interested in that's very volatile, or that changes a lot, and one example would be the circulation status information for an item in a library. If you are trying to build some type of union catalog system to support resource sharing, it could be very useful to know up to the minute or to the second whether something is available. And one very easy way to accomplish that is to basically send the searches out to the catalogs. There's other ways of accomplishing the same thing but that's one way of doing it.

Another type of situation where you may have to do this would be if the owner of the data doesn't want you to have it in your index. This is a little bit of a moving target, because I think there is an increasing realization amongst people who have content that it can be beneficial to provide their metadata or their data to people who want to build indexes. But there will still be people around who for one reason or another simply don't want to give you their data.

DAVID: But they do want to let people search so with federated, people can search, but there's no extra copy of their content on somebody else's site...

SEB: That's right! And here's another example that may come up...it may be that one of the sources that you're interested in searching is itself a super index, if you will, and it may be a scale of index that you may not be prepared to run a copy of. It may be that one of the sources you want to search is all of WorldCat with a billion bibliographic records or the commercial index that Serial Solutions is offering to

some of their customers. Those would be very challenging things to simply copy into an index on your own server.

DAVID: So does this get at one of the weaknesses of federated search, namely that, one imagines but you'll correct me, that if you're doing a federated search across thousands and thousands of sites that it can be relatively slow?

SEB: It can get challenging. I think what happens in particular is that as you add more and more sites that you want to search unless you are very careful in terms of how you put together your application, your whole stack of software, you can get stuck waiting for the slow sites, or the slow servers, to respond and that can make your whole system feel slow. Even worse if you really need to have everyone respond before you feel that you can deliver a complete response to your user, it's really unlikely that all of these sites will be available at any given time. So certainly if the goal is to be able to search in an absolute way everything, then there really isn't a substitute for having everything right in your hands and building an index on it.

DAVID: Can you explain the super node idea as a way of scaling up federated search?

SEB: So I think this is something that came up in our conversation. It was basically this notion that, and this has been sort of my practical experience, but I find that in trying to build a federated search system, and we spend a lot of time trying to play with how to make that perform as well as possible, and so as an exercise we try to look at where do we run into performance restrictions when we add more resources, be that memory or CPU or bandwidth, and in practical terms our software on a decent size machine might be able to search say a couple of hundred resources. Now that's an obvious limitation if you say that you want to search two thousand resources. So an approach that you could take and that some people, for all I know, may already be taking, is to have this notion of a hierarchical approach to the problem where you have these super nodes that will broadcast searches out to, say, a hundred resources each and then feed those results upwards through the hierarchy until it finally ends up at the end user. If you do that you will be able to broadcast searches to a very large number of resources.

DAVID: You are also interested in hybrid search systems?

SEB: Yes. When it comes right down to it, I think, that probably is...if you look at the general problem of providing users with access to information through the particular mechanism of searching which is the game that we sort of are playing, there is a debate, and we've just kind of talked through some of the issues, in terms of do you accomplish that by doing a federated search or do you accomplish it by building an index where you're gathering everything that you could want to search. And my basic notion is that neither one of those solutions in practical terms can really solve the entire problem. And if your goal is to have a technology that can solve the whole problem, it needs to use both types of techniques because there will be very volatile or very large sets of metadata that you can't practically ingest in your own index. On the other hand there are probably too many resources and too many of them that are running on too many different information systems with varying capabilities that is practical to do everything in a federated search. So my notion is that you want to be able to gather stuff together in an

index when it is practical and possible, and you want to be able to federate for the stuff where it's not practical or possible. And you want to try to do both of those things as well as you possibly can and you want to try to somehow get the results of both of those types of searches back to the user as a single nice friendly merged search results.

DAVID: The merging of results from a federated system can be a challenge, can't it?

SEB: Yes. There's a whole number of different issues that arise. Some of the really basic stuff is that you need to, on the fly, be normalizing things that come back to you in different character sets and different transfer syntaxes, different logical schemas. But you are also dealing with things like different conventions for representing the same thing. One type of source might talk about a book, as an example, with the title and the sub-title merged together in one field whereas another source will break them up into two fields. People will have different conventions for talking about what something is: is something a book? is it a web page? is it a journal article? or something else entirely? There are a number of different schemes you can use, and not a whole lot of standardization outside of the narrow library catalog space in terms of how you describe things like that.

DAVID: So it's the normal library semi-modified chaos when it comes to metadata, except with the federated search you have to resolve those differences in real time.

SEB: You have to do it in real time, and you have to do it very quickly, because if you want to try to present the user with a nice immersed result, you need to retrieve, generally speaking you probably will want to retrieve more data than you are actually showing to the user so that you can look at a larger set of data and try to make some determination of ranking and sorting and so on. For medium large result sets you may even try to retrieve the entire result sets, so everything has to happen very fast. And that's a huge part of the challenge.

DAVID: Which your company, Index Data, has some experience in, I assume?

SEB: Yes, so you could say that that's the specific area that we have chosen to play in over the years. And we've done it for 18 years now, and I wouldn't say by any means that we have solved the problem, but it's what we consider an interesting problem space and one that we've chosen to specialize in. You know, the one area is the on-the-fly processing of data and building some form of temporary result sets that you can then analyze and break down in various ways and display to the user. But also figuring out how you merge that data that comes back on the fly with the stuff that you are pulling out of the local index.

DAVID: And that's got to be hard for another reason that when we talked earlier you mentioned, which I find really fascinating, which is the difficulty of providing unified and ranked results that have been gathered from multiple sources.

SEB: That's right. One of the big benefits of having everything in one index is that you have a large set of data and most statistical relevance ranking algorithms like to compare the individual records in a result set against your query and then against the totality of records in the whole database. And by that they

form some notion of what constitutes an important record or not. When you are operating in a federated search environment and you don't have access to the contents of the whole database, you have to do without a lot of that information, and that makes it more difficult, and it also makes it probably impossible to accomplish a ranking that is as good as what you could do with a full index. One way that you can potentially get around that would be to basically encourage people who run, in particular, perhaps important information resources or large indexes to think about making APIs available to other people that provide access to some of that information. So in addition to providing say metadata associated with the record and maybe providing a simple numerical score, you might provide a vector of information in some normalized and standardized format that would allow a federated search system to better interleave ranked results from different sources.

DAVID: And that would help not just federated search systems but anybody who's doing a multi-source API query.

SEB: Exactly. So that comes down to looking at what are the actual needs of the end users that we are trying to serve in a community and is there a basis for trying to encourage information providers to make their stuff available in a form that lends itself better to what we're trying to do with it.

DAVID: So what are some of the weaknesses of a union search, of building a big index on a central server?

SEB: I would think there are a number of issues that pop up. But I think one of the biggest ones that we've encountered when we do it in practice is just the complexity of establishing a reliable feed of data and then keeping that healthy over time. To take an example if there's a source of data, some database provider or even a little library, in order to get them into your index you have to first speak to them and get their agreement to provide you with their data, unless you are simply going to crawl their website which is what Google would do, but because you can get richer data typically by speaking to them and actually getting an export of data from them, you typically want to speak to them. When you have agreement with them then you have to negotiate some exchange mechanism, including both a physical representation, a file format, a logical representation of data, and then a mechanism for getting that to you. Typically pretty much every time we've done this, there's a back-and-forth because it turns out that there's been misunderstandings on one side or the other, so there's a manual labor involved in debugging each of these feeds. And then having been established, depending on the technology, there may be a manual process on both sides to keep that data in your index synchronized, and that aspect can be a huge challenge because that has to survive over a long period of time, possibly through staff changes, infrastructure changes, and so on. So if you look at a very large union index made up of maybe hundreds, or even thousands of different sources, the cost of maintaining that can be really substantial.

DAVID: Whereas for federated all that you would need is to know what the particular API queries are for each of those sources?

SEB: Right. In some cases you'll have an API that wants to find and, once implemented, will tend not to change too often. In other cases there will be actual support for a standard space protocol for

information retrieval like Z39.50 or a descendant of it called SRU. When those are implemented it becomes fairly easy and fairly stable to hook something up. If none of those things are available, you end up having to, not to put too fine a word on it, screen scrape someone's website, and that can be a real hassle, and that's really where most of the cost and complexity of doing federated searching lies in building and maintaining gateways to those types of resources.

DAVID: So I will put in the plug that you are too decent to go for which is much of the work that your company does is open source and available but you do have a really pretty slick and useful tool for creating screenscrapers, and that is a proprietary tool but it demos real good. What's the name of it?

SEB: We call it the Connector Platform, and thank you very much for that. It is pretty fun, and it basically was something that we had to put together because we didn't have the infrastructure in terms of staff, manpower to really run a very large organization of programmers just to maintain those types of gateways. So we had to come up with a different solution and ended up building this thing that makes the whole process of both screenscraping but also gatewaying to people's proprietary APIs a lot more manageable, and even in some cases a little bit fun.

DAVID: So where do you see federated search going, either as a technology or within the ecosystem, so to speak?

SEB: I think it will have its place next to large indexes as simply a way that we enable access to resources when that is the best way to do it. So I think ultimately down the road the common approach for many people is going to be running a large index and then supplementing that with a federated search of resources that for one reason or another are not practical to get into that index, either because you don't have the wherewithal yourself to add them, or because you are buying that index from someone else, from a commercial vendor, and they don't have the wherewithal or the time to put the resource out there. There's a second step that we see developing right now with some of the folks that we work with where you're starting to see in the main of commercial services for libraries, there's a bit of a race in the marketplace at the moment amongst a few large vendors as to who can build the largest index of resources of interest to libraries. So that includes not only bibliographic material about books but various types of commercial collections of journal article citations. And we have seen more than a few cases now where a library or consortium is looking to purchase access to one of these big indexes, and I should say that they typically come fully formed with a nice-looking interface and all the trimmings, but they are finding that that large index doesn't cover all of their needs. They may be working in a specific area where there is another large index they need in order to have a complete information landscape. They may be living in a specific region where there are a number of resources particular to that area that just aren't in the big commercial index they are looking to buy. So essentially some of these folks are coming to us and saying we would like to do a federated search across two or three large indexes in order to present a complete information landscape to our patrons. So again I see federated searching is something that can supplement the large indexes, but I also really would like to move towards a place where we see the large indexes as utilities, so that not everybody that wants to build an information portal has to build their own large index, because as much as it has become a lot more manageable and a lot easier to do with current technology, it's still a lot of work. It will be better if it was easy for

different organizations to collaborate around building and maintaining these things. And that requires some form of federated searching, unless the indexes are small enough that you can just throw them around and make copies of them.

DAVID: So federated search is a way of allowing greater flexibility in pulling together across whatever set of indexes are out there that you care about?

SEB: Sure, and the exciting thing is that if you look at the big indexes and the types of technologies that they're based on, things like the SOLR Lucene indexing engine, they have probably the ability to fairly easily make available the type of richer interfaces that we were talking about a little bit ago in terms of supporting better access and better interleaving of ranked results from different sources. They also have APIs that might expose access to the internal faceting mechanisms of those systems so that you don't have to build your facets as part of your interface by counting records, so to speak, but you can get them directly from the source. So there are a lot of possibilities. If you gather data up into these larger super silos or large indexes, you can potentially expose interfaces that make it a lot easier to access them. By the same token, though, it's fair to say that you can also expose some APIs that would make it very easy to replicate them, like the OAI-PMH protocol for replicating a dataset. So it goes in both directions, but I think the bottom line, though, is that, I think, in all cases it would be useful for all of us to think again of an index as something you build, and once built, it's a resource. There is work involved in building it, and it will be useful if you can share it with other people in a controlled way.

DAVID: Sebastian, thank you very much!

SEB: Sure :-)