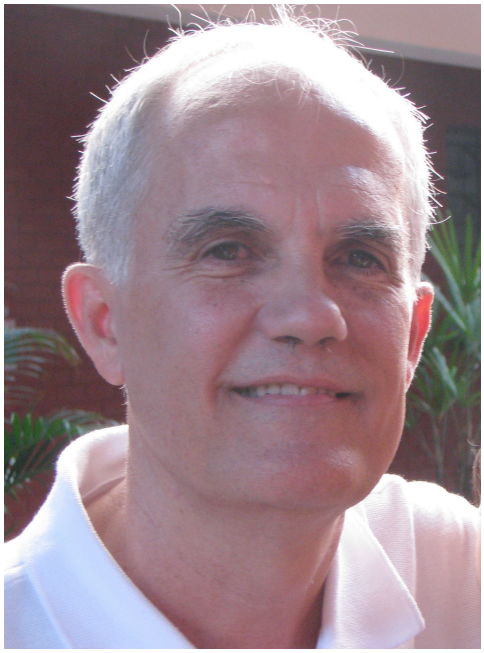
Alberto Laender Speaks Out

on Why Google Bought His Startup, How to Evaluate Graduate Program Quality, How to Do High-Impact Research in a Developing Country, How Hyperinflation Nurtured Brazil's Software Industry, and More

by Marianne Winslett



Alberto Laender
http://www.ufmg.dcc.br/~laender

Welcome to this installment of ACM SIGMOD Record's series of interviews with distinguished members of the database community. I'm Marianne Winslett, and today we are at the SIGMOD 2010 conference in Indianapolis. I have here with me Alberto Laender, who has been a professor at the Federal University of Minas Gerais in Brazil since 1975. His research interests include web information systems, digital libraries, and data modeling. Alberto is a member of SIGMOD's Advisory Board and SIGMOD's Jim Gray PhD Dissertation Award Committee. Alberto's PhD is from the University of East Anglia. So, Alberto, welcome!

Please tell us about your startup company, Akwan.

Akwan was created in February of 2000. We had developed a search engine for Brazil, which was called *TodoBR,* and we started the company. We used the technology to provide service for the community, for industry, for companies. And then in July of 2005, Google bought the company. This was the beginning of the Google research and development center in Latin America, the first one. In fact, I believe that Akwan was the second company that Google bought outside the United States. The first one was from Canada or Australia, I'm not sure about that; and the second one was Akwan.

How is search with Akwan different than the normal Google search?

At that time, we had what is called a vertical search engine. It was completely dedicated to the Brazilian web. Because of that, we were able to do much better than Google at that time in Brazil. We knew about the Brazilian web, and everything was customized for the Portuguese language, and also the search engine was very fast too. And so it was the favorite search engine for the Brazilian users at that time. Nowadays, Google is there, and they have a much larger infrastructure than what we had at that time, so they can do better. They also hired all our engineers, so all the technology that we developed is now being used by Google Brazil.

Many database researchers have been lured into industry by the promise of access to real data. Since Google Brazil must have incredible access to web data, why aren't you working for them?

That was part of the deal. When they bought the company, as I said, all the engineers stayed with the company. They were former students from our university. Berthier Ribeiro, who was the CEO at that time, was the only Akwan founder that stayed with Google. I don't know, maybe it is some kind of policy from Google that they buy the companies, and they keep just the engineers, so that after that, they can try new ways of managing things, and new directions. We don't actually have a kind of agreement with Google at the moment, other than perhaps we have been helping them to hire new people in Brazil – that is the only thing that we do with them.

So you are not using the data from them in your research?

No, not really. This is a pity, because it would be interesting to have some kind of data from Google Brazil to know a little bit more about the Brazilian web nowadays.

You have experience in measuring the quality of CS departments, an area where Brazil fares very well. One aspect of that task is measuring the output and impact of individual researchers. In the rest of science and engineering, people like to use the Thomson ISI impact factors. But you and your colleagues concluded that "a look at the Thomson ISI list of the top 250 researchers with more citations in the CS field shows that this list includes none of the ACM Turing Award winners in the last 10 years, clear evidence that this database does not cover the relevant publication venues in the field" [Assessing the Research and Education Quality of the Top Brazilian Computer Science Graduate Programs, *inroads SIGCSE Bulletin* **40**(2), June 2008]. *What do you think is the best way to measure impact, for computer science researchers?*

That is a difficult question. I should mention that we got this example from a presentation from a researcher from Europe who was looking at the ISI impact factor. He brought in data saying that among the top researchers, he couldn't find any of the Turing Award people there.

In Brazil, we have a Ministry of Education agency, CAPES it is called, which is responsible for evaluating all graduate programs in Brazil from all areas. They create what we call the *Qualis*, a ranking of journals and conferences for all areas. This ranking divides all the venues into different levels, and this is used to assess the research produced by the graduate programs. Because they try to do this in a general way, so that it could be used for all areas, they have been using the ISI impact factor for some time. In the case of computer science, using the ISI impact factor is a problem because it is difficult to access the impact factor for conferences, and conferences are very important for computer scientists.

Also, compared with other fields like chemistry or even biology, the number of journals in computer science is not so large as in other areas, so sometimes it is really difficult to have all those papers and all those venues correctly classified when taking just into consideration the impact factor.

I don't have an answer for what would be the best way of assessing all the graduate programs. Clearly some kind of impact factor is important, but the ISI factor could not be seen as the right answer for this kind of evaluation for computer science.

Didn't you come up with a new way based on DBLP?

Yes. We used DBLP to do a comparison between Brazilian programs and programs from North America (the United States and Canada), and from Europe. The idea was to show that the Brazilian computer science programs had a good standard compared with major programs outside Brazil. So we took data from DBLP and we used the Qualis system for classifying the papers and the articles from all the researchers in these programs. Then we did a vertical comparison, where we compared each computer science graduate program in Brazil with the ones from the United States and the ones from Europe. I think this makes more sense than just trying to give some kind of grade for a program. The result was that the Brazilian programs are not as good as the top programs in the United States and Canada, like MIT or Berkeley or Stanford or Toronto, but we are in the middle. We did quite well in comparison with other programs in North America and also with programs in Europe. So that was the idea, to provide an assessment about how computer science was doing in Brazil, compared with well-established programs outside Brazil. The results of this study appeared in the *SIGCSE Bulletin*, and are available on line at http://portal.acm.org/citation.cfm?doid=1383602.1383654.

Why does Brazil produce so many database researchers?

That is a good question. I think I have two possibilities for answering this question. The first one is that when the graduate programs started in Brazil, databases was something very attractive, because it was a new area at the time. This was about 30 or 35 years ago. The relational model was just getting a lot of attention then. This attracted a lot of people. This I think was one point.

The second possibility was that the first graduate program in computer science in Brazil was the one from the Pontifícia Universidade Católica (PUC) in Rio. And there, Prof. Antonio Furtado, who is well known in the database community, started a very nice group and was responsible for graduating a lot of PhDs at that time. His former students started other database groups around Brazil. I think this was the main reason why the database community there was larger than the other ones.

As a consequence of this, the Brazilian Database Symposium was one of the first Brazilian computer science symposiums in Brazil. We are celebrating this year the 25th anniversary of SBBD, as we call it. SBBD nowadays is recognized as a very strong symposium in Brazil, and also it is seen outside Brazil as one of our major conferences.

If I look at the 2009 photo of SBBD attendees, I see so many female researchers. Why is that?

I really don't know, but this is true. The female community in the database area in Brazil is very large. It might be because when the database community started in Brazil, it was very common to have girls in the computer science undergraduate programs in Brazil. From some time, almost 50% of the students were female. But, nowadays it is not that high. I think it is the same problem here in the United States, that the number of women in the computer field has decreased very much. And it is like a tradition, we still have a lot of females working on databases in Brazil.

In the 1990s, Ricardo Baeza-Yates wrote that "except in the dubious domain of providing bodyshopping labor, it is proving very difficult for embryonic software industries in [lesser-developed countries] to be competitive in mature markets." Do you agree?

I can say that I don't agree, but I think that Ricardo is right in some sense. I think that in all developing countries, you can always find a way to do some kind of good research in computer science, and also transform these research results into some prototypes, and even have some startups as well.

In the case of Brazil, for instance, we have a very interesting situation now because the bank industry in Brazil has a very sophisticated computer system. The main reason for that was the fact that years ago we had a very high inflation in Brazil, and the banks had to be very efficient to process all the transactions.

Why does hyperinflation mean that banks have to process transactions fast?

With high inflation like Brazil had, if you just deposit a check, you cannot wait for one week to have your check cashed. You need the money almost immediately, otherwise the inflation will just, how can I say… your money can disappear in a few days, or at least lose a lot of its value. Because of this, all the bank financial systems had to be very efficient, and everything was automated.

The bank software that was developed in Brazil is nowadays being used in many other countries, in Europe, in Asia, and all over the world. And because the banking system and financial transfers had to be very secure, network security also is an area where Brazil has many companies who built software that is used outside Brazil today.

So the unique national circumstances of that hyperinflation became an opportunity for the local industry in Brazil. There is a saying about that: "If life gives you lemons, make lemonade."

Yes, exactly. That is the point.

What approach do you recommend for doing high-impact research while living in a developing country?

It is difficult to give a single answer for that question. I think that people have to look for good problems that are very interesting for their students, and also that can be used for many other people. For instance, I think the web is a great opportunity because the web is a very democratic environment. If you do some kind of research on the web in Brazil, or in Asia, or in some country in Africa, or here in the United States, what you do has to work for the whole web. If it works for the whole web, you have the opportunity to show that those ideas that you have are good for some kinds of applications, and that you can construct a useful system from there. You can have the opportunity to have some products, or some services, with the web as a basis. So you have a very nice environment to check and to test all the applications that you have developed. I think that the point is to find good problems where you can construct a prototype for their solution, based on the web. I think the web is a great opportunity for doing research in computer science in developing countries.

Can you give us examples of topics that maybe weren't so great to pursue in a developing country, where it is harder to have an impact?

For instance, trying to do something in computer architecture is almost impossible in a developing country because you need a whole industry behind you so that you can have impact in that. On the other hand, if you have a new kind of processor, you can do some research on how to use that processor for a specific application or how to improve, for instance, query processing in database systems; that you can do nicely. But trying to develop a whole new database management system is something that is hard work, and is very difficult to do in a developing country.

Tell us about your project where you are analyzing posts on Twitter and Facebook.

This is a project from our group, which is part of the National Institute of Science and Technology for the Web. It is a new kind of project that has been approved by the Brazilian government, and we are one of the four national institutes related to computer science in Brazil. This year is the presidential election in Brazil, and so we developed a whole framework to try to follow what is going on on the web in some specific subject. For instance, we have the elections in Brazil, we have the soccer World Cup in South Africa, and these are very sensitive subjects in Brazil. And so what we did was to follow Twitter, Facebook, all the things that happen on the web every day, and try to put all this data together so that the users could see what is going on in terms of the election, who is the candidate that has been exposed in the media, how the media is talking about the candidates, what was the discussion in terms of the major points for the election. Now we have the soccer World Cup in South Africa, and so we are doing more or less the same, following the players, the teams. Giving to the user an idea what is going on on the web about these subjects is very important, at least in Brazil. Soccer and elections are two subjects that are very important for Brazilians.

It sounds a little bit like another area for a startup company.

Sure, we do have a chance to do that. We have shown some results of this project to media companies, and they want to do the same thing for the brands of their clients. They want to show a customer how his company is seen on the web, and so we can use the same approach to help them. In fact, most of these media companies already do this, but they do it manually. They go to the web and try to collect data about a specific brand, or a specific company, and they have a lot of people doing this manually. Using some techniques from data mining, from classification, we can do at least a first step automatically, and give them a lot of results. Then they can use the people that they have behind the scene to check whether the data we collect from the web is okay. For them, it's much better doing it this way than trying to have people look at all the pages on the web, all the sites, to determine this feeling about the client's brand.

I know that other people in the database community are also working on sentiment analysis and that sort of thing. Is it different in Brazil, just like your search engine was different for Brazil? Is it customized in some way?

I don't think so, I don't think that it is different. It is interesting, for instance, that Brazil has the largest Orkut community in the world. Nobody knows why. Google is a major player there now, but when Orkut appeared suddenly in Brazil, all teenagers had to have their account in Orkut, and they use that to communicate among themselves. Brazilians like all the facilities available on the web today, and they are attracted to the things that happen on the web. People nowadays use all the social networks, and it is almost a must for the young people to have an account on Facebook or Orkut or any other social network.

Sounds like the US.

Yes.

Do you have any words of advice for fledgling or midcareer database researchers or practitioners?

The researcher must try to find a good research topic, and must learn about the work on that research topic that is going on in the world. They should try to cooperate with people who already have some experience on that topic. But particularly, in choosing the problem, it must be something you like. If you find something interesting, go after it. I think that is the main point that I would say to a young researcher, or people who are trying to start some kind of research in computer science in general. I think computer science is a fascinating area, and we have a lot of things to do in computer science.

Among all your past research, what is your favorite piece of work?

There are two works that I like best. The first one was when I started to work on traditional database modeling, and the second was when I used traditional database modeling techniques to help to develop a data extraction tool for web data.

I started my research career working on conceptual modeling. At that time the entity-relationship model was something that everybody used. I did a very nice piece of work together with my colleague from PUC Rio, Marco Casanova, and a PhD student of mine at the time. We developed some techniques for mapping from the entity-relationship model to the relational model, in such a way that you could correctly represent everything that was expressed in that ER diagram. It was interesting because it was a kind of logical database design method. The results that we got at the time were very useful in practical situations. We even implemented some tools, which is something that I like very much to do. When I returned from my PhD, this was my first research project, so it was very important for me.

Much later on, I was doing quite traditional database research, and I wanted to change. The web already existed, but we didn't know very much about the web. I had the opportunity to spend a few weeks at HP Labs in Palo Alto. There I worked with Moshe Zloof and started trying to do something somewhat related to Query By Example. Moshe had a project that was called the Picture Project, where they were trying to use the same paradigm as Query By Example, but also for creating interfaces and creating and developing software by example. On the web at that time, you found a lot of data, basically HTML pages from many different sources. I had the idea that we could extract that data so that we could store it in a database and develop some traditional applications. I had the idea to use the paradigm of programming by example for this task, and we created DEByE, which is Data Extraction By Example. This became the major topic of a PhD student of mine. We developed a tool that was one of the first data extraction tools that could extract data from web pages that had some kind of internal structure. Most of the tools at that time were able to extract only tuples, and we were able to extract data with a hierarchical structure, so it was a very powerful data extraction tool at that time. Our results appeared in a survey that we wrote for *SIGMOD Record* on data extraction [A. Laender, B. Ribeiro-Neto, A. Soares da Silva, J. Teixeira, A Brief Survey of Web Data Extraction Tools, *SIGMOD Record* **31**(2), June 2002; http://www.sigmod.org/publications/sigmod-record/0206/laender-survey.pdf]. We had another paper on that topic in *Data and Knowledge Engineering* as well [A. Laender, B. Ribeiro-Neto, A. Soares da Silva, DEByE - Data Extraction By Example, *DKE* 40(2), 2002, http://www.sciencedirect.com/science/article/pii/S0169023X01000477].

Later on, we had a paper in the WWW conference [D. de Castro Reis, P. Golgher, A. Soares da Silva, A. Laender, Automatic web news extraction using tree edit distance, WWW 2004], for which we developed another tool for extraction of news from the web. We used this extraction technique at Akwan for developing clipping systems for news stories, which extracted the headlines of news from the web. Using that, you could provide for companies a summary of the news from all the newspapers and the web sites that we had at that time. This paper describes a very interesting technique for extracting a particular kind of data from the web, news. You have to recognize exactly all the specific parts of a web page that you want to extract.

If you magically had enough extra time to do one additional thing at work that you are not doing now, what would it be?

We always want to do a lot of new things, but I would like to have more time to stay with the students at the labs, and maybe do some programming. I liked very much to write programs when I was doing my PhD studies. After you come back and start your own group, you have to manage all the students and you have to write project proposals to get some grants, and the end result is that you are very far from the activities in the labs. So I would like to have more time to be there. I had some opportunity for doing

that when we started our company, because I stayed there for almost 6 months working with the engineers on some projects, and this was very nice. But I would like to do this again with the students. Maybe when I retire, I can come back to the lab to do some programming with them.

If you could change one thing about yourself as a computer science researcher, what would it be?

This is a difficult question. After so many years, it is difficult to see what I should have done.

I decided to work on databases. I think that if I had enough time, I would like to go to see a little bit more about other areas, like computer networks, for instance. With the web today, I would like to be able to understand a little bit more about the whole aspects of the network, the physical aspects of the network, but it is almost impossible to do this with the responsibilities that I have now. I think that sometimes maybe when you start doing your research, you dedicate too much time only to that subject, and there are other areas that are so interesting. But it is difficult to look at more than one area at the same time.

Thank you very much for talking with me today.

It was my pleasure.