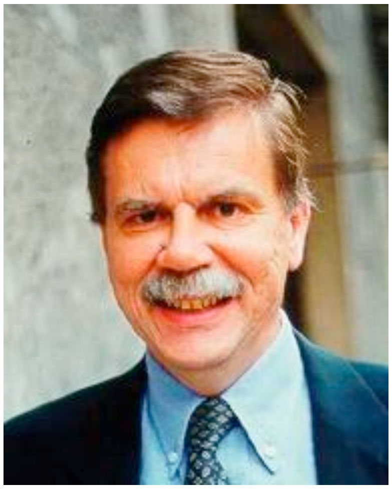
Stefano Ceri Speaks Out
on Many-Book Researchers and One-Startup Researchers, Web Modeling, the Vanishing US-Europe Research Gap, the Semantic Web Services Train, and More

by Marianne Winslett



http://www.elet.polimi.it/upload/ceri/

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 L3S Research Institute in Hannover, Germany, where I am spending the fall of 2006. I have here with me Stefano Ceri, who is a Professor of Information Engineering at the Politecnico di Milano. Stefano's research interests include distributed, deductive, active, and object-oriented databases, and XML query languages; his recent work is on the design of Web applications. He is a co-inventor of WebML (Web Modeling Language), and has a startup company that is commercializing WebRatio, a product based on WebML. Stefano was a founder of EDBT and is still a member of the EDBT Endowment, and he has been a member of the VLDB Endowment for twelve years. He was an editor of ACM Transactions on Database Systems *and* IEEE Transactions on Software Engineering*. He is a coauthor of 9 books, and his PhD is from the Politecnico di Milano. So, Stefano, welcome!*

Thank you very much.

Most researchers put most of their energy into conference papers. What led you to write nine books in 18 years?

I like first to do research in an area, and of course write conference or journal papers. Then, when I feel that I understand enough about the area, I like to write a book about it. That is also a way to get away from that area, because writing a book is normally the last thing I do on a particular topic before moving to another area. I did that for active databases, for deductive databases, for distributed databases, and for conceptual database modeling. This did not happen with the last book that I wrote, about web design, because I am now doing a startup on this topic. So that is the exception. I remember discussing this once with Jeff Ullman, who said that there are zero-book researchers, one-book researchers, and many-book researchers. We both belong to the many-book category, although Jeff has written many more books than I did.

You also have a lot of journal papers. Do you favor journal papers over conference papers?

Conference papers are not that easy with my particular style of research. My research is mostly on modeling, on trying to understand users' requirements and turn them into data management or web applications. To some extent conference papers must adopt a standard kind of syntax, the well known Wisconsin model: you first define a problem, then define the solution, then quantify how much better the solution is relative to the previous solutions. If you get a 30% performance advantage then you are fine, otherwise you don't even submit the paper. This kind of format gives a hard time to whoever has no quantitative measures to provide. I think this is a mistake that our community is doing, to leave away from the conferences the contributions that cannot have such a quantitative description and analysis.

Then with your work on modeling, how do you prove that your modeling approach is better?

It is very hard to prove that a given modeling approach is better. I do think that being able to provide the best models, the best tools, the best environments, and so on, is an important factor for the success of our field. But these things are hard to measure. For instance, people in the software engineering industry measure *function points*, which are quantifiable. But it is more difficult to quantify results for something which is a model. That is one of the reasons why we turn to journal papers more than conference papers.

Another factor is that the journal review cycle is a process where you have a discussion with reviewers. After a couple of iterations you have explained your point of view, and the reviewers can understand that what you are doing makes sense. In the conference reviewing process, sometimes the outcome is more a matter of taste, or luck, or being assigned a reviewer that doesn't understand the approach. So I think the reviewing process is more under control when you publish in a journal.

Of course the community likes a lot of conference papers, so we do like to write them. Recently I had papers in the WWW conferences but these are less visible in the database community than SIGMOD or VLDB.

Is it true that you got your PhD before your Master's degree?

This is a funny story. When I got my degree in the Politecnico, there was not yet a PhD program. I got the maximum degree I could get, which was a doctorate degree, but it was not equivalent to a PhD. The degree was in electrical engineering, and I did not feel that I had enough background in computer science. So after being a researcher for some years and publishing my papers in VLDB and so on, I went as a Master's student to Stanford University. That was 1981, and in retrospect, I had a very nice time. I had the chance to meet database people such as Jeff Ullman and Gio Wiederhold, and also other people like John Hennessy, Sue Owicki, Bob Floyd, and gurus like John McCarthy. Hector Garcia-Molina finished his PhD and Don Knuth gave his last lecture the year when I was there. All these people were sitting around in this small computer science department and you could breathe their presence. This was really exciting; it was really a good time for being at Stanford.

In 1990 you wrote a paper called "Object orientation and logic programming for databases: a season's flirt or long-term marriage?" Which has it turned out to be?

Probably a seasonal flirt with some children that were not wanted from the very beginning! The object oriented system promoters wanted to have more success and more visibility than they have had. But object relational databases are now very important. Also the marriage that now takes place between relational implementations and object oriented languages such as Java is very important. So the object oriented approach has become more and more influential, although not along the lines where the manifesto of object oriented databases [Atkinson et al., DOOD89] expected it to be.

Rules didn't have as much success as the object-oriented approach, but they are also very important. When I teach my class, I always teach deductive and active rules as two important concepts for data management. This gives students the idea that the database is not just a box for storing content; it also has all kinds of knowledge about this content. To some extent this gives the field a higher status that the students appreciate. And then there *are* many applications that use rules.

You've also worked on supporting workflows, as have quite a few other members of the database community. What impact has workflow research had on how companies conduct their business?

Companies do conduct their business using process modeling. They model their work processes by means of BPMN, BPEL, and so on. From our experience with workflow modeling, we have learned that business process modeling is important. True, the technology of workflow tools is not very much used. But the idea of modeling a process is picking up very solidly, and has an impact in industry.

Why don't companies use workflow tools also---what is the missing ingredient?

They don't use workflow tools, but they are very serious about process modeling. For instance, we are negotiating a contract with a big Italian company that would like to standardize how they go from data and process requirements down to concrete software code.

So that company's process is the process of creating software.

Yes, the software design process. I think there is a need to model this process and then to use tools to produce software from models. This is not done too much by computer scientists; it is more done by industry people, but the models that they use are not supposed to be turned into implementations. They are not expressive enough.

Should computer science professors worry about technology transfer of their research?

They should, they definitely should! In my department, if I consider how much money we get for research, only 5% comes from internal grants (from our university), and 95% comes from external contracts. So we have to get contracts, but for them we need to be able to solve real-world problems, hence we need to do some kind of technology transfer.

Starting from a real-world problem is always good. When you develop a solution, you can think about whether your solution is good enough to be generalized. So, there is still space for invention, abstraction, and so on, even if you start from a very concrete problem. If you don't start from a real-world problem, you risk creating an abstraction that has no relation whatsoever with reality. If you start from a real-world problem, it becomes much easier to abstract the right concepts.

Another question is how much real-world problems are valued in the computer science community. Some computer scientists don't really appreciate the advantage and the beauty of creating a solution to a concrete problem. They still see a separation between theory and practice, even though theory can be very much applied to concrete problems. They think about their problem, and then think about their solution, and then they are done. Researchers should find an external reason, a real-world problem, as motivation for starting their work; after that point, the work may become very abstract, but at least it has some concrete foundation.

You are the chairman of LaureaOnLine, a "fully online curriculum in computer engineering." From this experience, can you tell us what the trickiest issues are in on-line education, and whether it will replace face-to-face college education?

Let us start from the last question. Replacing face-to-face classrooms---no, that is not what we want to do. LaureaOnLine is a full bachelor's curriculum, so we offer 28 courses of 5 credits each that can be composed into a curriculum. Politecnico has really invested efforts in this program and done it in a consistent way. Our graduating students don't have a degree that says "online," it only says "computer science." Our online program has exactly the same quality and seriousness in the teaching as our on-campus program.

The students for the online program are typically working people or people who live far away from Milano. So our goal is to address a different student population. We have things such as virtual classes, which use synchronous communication technology; students meet at times which are good for them, like in the evening or during the weekends. They can do co-browsing, co-editing of shared resources. They can do projects. There is a tutor who can explain the class material. This online communication actually sometimes turns out to be even more alive than the communication you get in the classroom. The online students form a community, which is very nice. I really like the kind of relationship that is established among online students.

The online students have to come to the Como campus to take exams. They come to campus for one very intensive week twice a year. That is similar to the Master's system in the United States, I think, where students have a final week of exams. In Italy, however, exams are scattered all around the calendar year, so that's an exceptional situation.

So you are saying the average student is taking more than one course at the same time, even though they are working?

Yes, they normally take about half of the courses that an on-campus student would take. An on-campus student would do the curriculum in 3 years, so an online student might take 5-6 years.

How many students are enrolled?

We have about 370 online students total.

There are lots of other uses for the online courses. For instance, we have Master's programs whose students take the most advanced online courses; they can be taken as well by new graduate students who have "debits", i.e., exams that they didn't do in their past curriculum.

What have you found to be the hardest issue you have faced in making this program work?

To let the people know that the program exists, and to communicate well how the program works.

Do you see any new data management challenges that the database research community should be aware of?

I think the database community should be more aware of the problems of the web, but that's obvious because I am doing a lot of research on the web. I think the database community has somehow missed some trains that have left the station. It is true that if you look at the major companies such as Google or Amazon, they employ DB researchers who know perfectly well our technology. But we are about to go to another level of complexity in the web, the so-called semantic web services, and this is a train that the database community should not miss. The ability to describe services in a massively scalable way, the ability to search for the right service, the ability to reason about what is the best "opportunity" for a user request to be serviced on the web---it is a problem that we should be more aware and more conscious of. Otherwise, other people will solve it. I think that these semantic web service facilities will take shape in the next 5-10 years.

The issues that you mentioned are things that professors in my department are working on. So when you said that what Google is doing is not represented in the universities, do you mean that we should have a course that talks about those sorts of issues?

No, that would be too extreme as a position. But for instance, take research on personalization: it requires monitoring and extracting knowledge from the domain and from the user. Do we want to solve this problem, or do we want other communities to do it instead? More generally, I think that we don't have at the moment a critical mass of database research which is oriented toward solving the massive problems that you may have if you view the web as a large database. That may be a wrong perspective ---maybe everybody is working on those problems---but that's my feeling.

You've spent dozens of summers at Stanford. What do you see as the main differences in the US and European approaches to research?

There are lots of differences. First of all, resources. Second, the closeness to a market which is very responsive. In the US, as soon as you do research that has some potential, immediately people come to you and talk to you about what could be done with your research. In Europe, it goes the other way around. We have to dig out the contacts. Technology transfer is very different in the US and in Europe.

In terms of research, sometimes I think that in the US, you tend to go more deeply into a narrower field when you do your PhD thesis. You become the world's expert in that particular field. But if you invest to go in depth, then maybe you don't invest to go in breadth. So that might be a difference between Europe and the US, in that we go less in depth and more in breadth.

You mentioned that I have been going to Stanford for 20 or so years. When I was going to Stanford the first time, I was feeling that there was a big gap between the US and Europe, or better say Italy---a 5 year gap, which I was closing by crossing the ocean. Now the internet makes the gap much shorter. Now I can live in Europe and read instantly what is happening in the United States, or in any other country. As a matter of fact, I didn't go to Stanford the last 2-3 years because I felt that I was not disconnected, I didn't have a gap to catch up. I felt that what we were doing was state-of-the-art; we could look it up on the web, and could have discussions with our colleagues over the internet. When we do discussion with colleagues in the States, we use all kinds of video communication technology, co-browsing and so on. And it becomes very effective. It is like being in the same place.

The European funding agencies seem to like big projects. How can you do research in such big teams?

That is another difference between the US and Europe. We have Information Communication Technology (ICT) projects (formerly called Esprit projects), each of which may have on the order of 4-

5 (for a small size project) up to 20 (for a big size project) partners, from many different countries. There have been six rounds of EU funding, and the seventh is coming up.

My group has had a project in each of these rounds of funding, and it has been probably over 70% of our funding. So I cannot complain. We have invented WebML – our Web Modeling Language - through one of these projects, and then we have developed WebML further through two other EU projects. So the EU funding approach has been very effective for my research (and for my start-up). I have not had a bad experience with this approach to funding.

Of course it is very difficult to set up the project; you spend a lot of time finding the right consortium. My experience is that if you write a good proposal then normally it is accepted. Sometimes people put together a consortium based upon what is "called for" in programs (and not upon their existing research links), and then these projects are much less effective. The EU approach is much more complicated than the US model where you apply to a funding agency to get your own pot of money, but the cooperation might be healthy sometimes.

The Lowell and Asilomar reports on future directions in database research [SIGMOD Record, Dec. 98; SIGMOD 2003]---what do you think about such reports and their impact on database research?

You mention Lowell and Asilomar because I was there both times. It was a very interesting experience. I remember having a very nice discussion with Mike Stonebraker, who was telling about how he sold Illustra to Informix. I don't remember exactly the details, but it was interesting to hear this from the front lines, so to speak.

At the Asilomar workshop ten years ago, I remember saying that three things are important in databases: semantics, semantics, and semantics---following in the style of Bruce Lindsay, whose famous words were, "There are three important things in databases: performance, performance, performance." I think that semantics is getting more and more important for the database community.

I have two students here visiting L3S with me, so I asked them whether they have ever heard about these reports. Their answer was, "Yeah, but we don't really remember what the reports said." So maybe these reports are less relevant to students than to people who are already established researchers, where the reports can help to build the field's consciousness.

Tell us about your startup company.

We did a startup using a patent that we have written while being employed by the Politecnico. So the Politecnico owns part of the startup. I co-founded it together with Piero Fraternali, a professor at Politecnico (and a good friend), and also together with three students, who are now taking the lead at the company. We didn't use any venture capital. To some extent, this may mark a big difference between startups in Europe and US. We can steadily grow, but it is much harder to have a big boom. Now we have a very good product, with many customers; the company is doing well. It has been very interesting for me, being a professor, to meet this totally different world, where the clients are always right, and you have to meet their needs.

Can you say a couple of words about what the products do?

The product is WebRatio, a tool that enables you to design web applications. You design the application by modeling---of course, that's my specialty!---so you have a model for the data, and then a model for what we call the *hypertext*, which is the description of the web interface, and then a model for the presentation, the look and feel. These models are orthogonal. Once you finish with the modeling, you just run code generators to create the software, so you never program. And the software is platform independent, so it works with arbitrary databases, arbitrary web servers, arbitrary environments. It is an application generator for the web. It is very sophisticated, we support web services, processes, rich internet interfaces, and so on.

We participated in the Semantic Web Service (SWS) Challenge, a competition to model B2B solutions for the semantic web in June 2006, where we used the WebRatio product as it was at the time. Ours turned out to be the most complete solution. I also did get a prize for SWS research, an IBM Faculty Award. So, the modeling of semantic web services through this model-oriented WebRatio tool is probably our next step. It looks to be very promising.

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

Look for an existing problem, then try to generalize this problem and come up with solutions. Then at this point you have both a practical problem on the one hand, and a generic solution that you can publish as well.

From among your past papers, do you have a favorite piece of work?

Actually, there are many. There are the summertime papers with Jennifer Widom, who I visited at IBM every summer while I was at Stanford. We had a VLDB paper together every year for four years. In my life, I have had a lot of coauthors who are wonderful people: Giuseppe Pelagatti, Georg Gottlob, Gio Wiederhold, Sham Navathe, Letizia Tanca, Ioana Manolescu. Piero Fraternali, Stefano Paraboschi, and I have been a "prolific trio", I remember seeing that in the *SIGMOD Record*. And there is a piece of work that is coming out which I like a lot on data mining, in the March 2007 issue of TODS.

Data mining? Now that's a new topic for you.

Yes, that's a new topic.

So what have you done in data mining?

We have defined a new data mining pattern. We call it pseudo-constraints. The idea is that you would like to find in the database things that are almost but not quite constraints because they do have some exceptions. Then you can express this constraint and find its violations. These are the interesting things that you mine from the database.

The violations or the constraints?

Both. The violation is our instance, and the constraint is our rule. If you combine the knowledge of these two things, then you can understand a lot about the underlying domain. For instance, a pseudo-constraint may say that students attend a given course's lectures if they are enrolled in the course; exceptions are "interesting cases", maybe errors to be fixed, maybe truly interested students from outside the university. As an extreme case, we discovered some weird things in public data about bank transactions used in data mining contexts.

What was the constraint being violated?

The constraint says that whenever two transactions are linked to a certain third party, they are issued by persons with the same birthday and sex but not the same name. We suspected them to be the same person. You can read about the details in the TODS paper.

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

I would try to combine my hobbies with my work---for instance, music. I think a lot can be done there. For instance, when I listen to music, I like to also read the music's full score, and I would like to see multimedia tools where you can read the score, listen to the music, see the performance, point to a line of the score and listen to a specific instrument. There is a lot to do in this kind of making music more accessible for the careful listener. That would be an interesting and challenging project.

There is work on music information retrieval, e.g., query by humming…

Yes, but I haven't yet seen in the shops the things I would like to buy.

Maybe your next startup company.

I don't know that I will have another startup company.

Maybe Jeff Ullman would say that there are zero-startup people, one-startup people, and many-startup people.

And I am a one-startup person!

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

I met my wife, Teresa, in Stanford, during the year that I finished my Master's. That was really a very good year. Then I had to decide whether to go back to Italy or to stay in the States. We had offers for jobs in both places, and had to choose what to do. We decided it was better to spend two months in the States each year and live in Italy the rest of the time, rather than spending vacations in Italy and the rest living in the States. That was a difficult choice. If I had made a different decision, it would have changed a lot in my career because then I would be closer to industry and to "opportunities". But on the other hand, I think I won't complain. It was nice to be where I was and to teach the students that I taught. So I'm not unhappy about my choice.

And you haven't missed out on the startup company boom either.

Well, we missed the boom, we came a bit later, when the startup market was already declining. That was better because if you hit the boom and then you are in the middle of the bust, it's a nightmare. I know many people who lived the "boom experience" badly. Of course, I also know many people who did very very well.

Thank you very much for talking with me today.

Thank you.