# *Sihem Amer-Yahia Speaks Out on Social Computing and DEI*

**H. V. Jagadish and Vanessa Braganholo**

## Sihem Amer-Yahia

https://lig-membres.imag.fr/amery/

*Welcome to ACM SIGMOD Record's series of interviews with distinguished members of the database community. I'm H. V. Jagadish, Professor of Computer Science at the University of Michigan. Sihem Amer-Yahia is my guest today. She is a Silver Medal Research Director at the French National Center for Scientific Research (CNRS) and Deputy Director of the Laboratoire d'Informatique de Grenoble, one of the largest research labs in Computer Science in France, with CNRS and INRIA Researchers and University Professors. She has won many awards, including the 2024 IEEE TCDE Impact Award, the ACM SIGMOD Contributions Award, and the VLDB Women in Database Research Award. Welcome, Sihem!*

Thank you, Jag, for the invitation and for the introduction. And let me also thank you for giving me a chance in your team as a postdoc at AT&T Labs in 1999. That was the start of the first chapter of my work life.

*You have made significant contributions in many areas, but if I had to pick one, I would probably name social computing in the context of data management. Can you tell us a little about how you came to this topic and what work you're most proud of in this area?*

In my career, I went from core data management questions, such as query processing for structured and unstructured data, to exploiting subjective human-generated data. My transition to social computing started when I joined Yahoo! Research in 2006. It was a time when the Web 2.0 was burgeoning. Web application owners understood the need for social interactions to drive traffic to their site, and pure social network developers understood the value of content. I quickly understood the importance of the social nature of data produced by humans, and I became convinced that to build applications where humans interact effectively with each other and with data, it was necessary to think of data models that capture human factors and behavior. In 2009, I wrote a CIDR paper with Cong Yu and Laks Lakshmanan titled "SocialScope: Enabling Information Discovery on Social Content Sites," where we introduced a graph data model, an algebra to manipulate data about people and their interactions, and a system architecture to build applications on the social Web.

***The amount of effort our community has deployed for all sorts of questions around XML storage, query design, processing algorithms, optimization, and later on, XML Full Text Search is tremendous. Even though that did not "make it" in the same way relational databases made it, I believe it had a wider and lasting impact on us as a research community.***

At that time, the work that was going on in social computing was very much observational. I went to conferences such as CSCW and ICWSM, and I met social scientists and psychologists. I became aware of the fact that there were so many theories in the social sciences that could be verified on the social Web. The hard question was which of those theories mattered and how to distil them into questions that mattered to me as a database researcher.

The work I'm most proud of in social computing is relatively recent because some of those theories only started to make it into my own research 10 years later, in 2017. It happened in the work of my student, Julien Pilourdault, where we examined the impact of human factors on designing recommendation algorithms on online crowdsourcing and labor markets. We read a lot about theories from the Psychology of Work that date back to the 70's, where they solved everything about people at work and their motivation, and we used that to formalize intrinsic and extrinsic motivation factors. That helped us better understand how to design adaptive algorithms that observe people as they complete tasks and capture their motivation to feed it into the logic of recommendations. That work required a fair amount of engineering, too, and led to a collaboration with Atsuyuki Morishima at the University of Tsukuba to build Crowd4U, an academic crowdsourcing platform. In retrospect, we only addressed the tip of the iceberg.

*Well, that's a lot, though!*

*You did some really impactful work on XML early in your career, including major contributions to XPath and a highly cited paper on tree patterns. At that time, we all thought XML was going to take over the universe. That hasn't quite happened, though there is a very solid niche for XML today. What is your opinion about that?*

The amount of effort our community has deployed for all sorts of questions around XML storage, query design, processing algorithms, optimization, and later on, XML Full Text Search is tremendous. Even though that did not "make it" in the same way relational databases made it, I believe it had a wider and lasting impact on us as a research community. The work of XML in the database (DB) community initiated a movement in our community. It made us rethink the fundamentals of query processing. It made us relevant to the Information Retrieval (IR) community and to Web standards. In 2005, when I was still at AT&T, I moderated a panel in SIGMOD on "DBs and IR: Rethinking the Great Divide". We debated the question of rethinking data management system architectures to merge DB and IR technologies. I remember being lost in translation when trying to bridge the gap between Boolean queries and the need for ranked retrieval for

XML full-text queries. That confusion is so much clearer today.

*Given how things turned out for XML, how do you feel about your own work and contributions?*

My work on XML was a pivotal moment in my career. I started collaborating with various researchers and practitioners. In 2003, with Pat Case at the US Library of Congress, we wrote a W3C recommendation document where we designed use cases for XML full-text search, based on how Pat and her colleagues in the library accessed documents on the Web. Pat taught me that to work with people from other disciplines, we needed to learn to speak the same language, so to speak, and align our goals. With Jayavel Shanmugasundaram, we added full-text search primitives to XQuery and XPath, and our language, published in VLDB 2005, was integrated into a 2011 W3C recommendation.

While at Yahoo!, I had a chance to work with great IR experts, Ricardo Baeza Yates and Mounia Lalmas. XML allowed us to do a lot of work together. We published papers and gave tutorials at SIGIR and VLDB. It's interesting to see how the topics of our collaborations have evolved over time. It went from XML languages and the INEX (The INitiative for the Evaluation of XML retrieval)[1], to questions that were more fundamental, about how to integrate information retrieval and database techniques to solve XML retrieval questions[2], and that led to questions around accessing data on the Web[3].

My work on DB/IR integration culminated with a panel at VLDB 2007 with Alon Halevy and a SIGMOD Record paper where each panelist defended their statement: Alon defended the idea that the Web 2.0 is about helping the masses manage heterogeneous datasets collaboratively. Gerhard Weikum promoted the fact that the Web 2.0 is about content-production democracy and a data-quality crisis. Volker Markl and Donald Kossmann focused on how one could use database expertise to define mashups declaratively, and AnHai Doan outlined pressing database questions in the Web 2.0. On my part, I talked a lot about how to leverage social ties to find the right content to serve to the right user. And that had a long-lasting impact on the way I designed recommendation algorithms that made use of social behavior and social ties. Two years later, in 2010, I found myself working on data management questions for human-centric Web applications.

Crowdsourcing became a central topic in my work after I sat on a SIGMOD 2010 panel moderated by Michael Franklin on Crowds, clouds, and algorithms: exploring the human side of big data applications. Ten years later, in 2019, I co-organized a Shonan workshop titled "Imagine all the People and AI in the Future of Work." So, in hindsight, working on XML got me closer to people.

*In all of this success, I assume there were ups and downs. We all have written papers that we feel were not appreciated enough. Is there any work you would like to talk about that didn't receive the attention it deserved?*

In a way, my biggest failure is my greatest success. My most cited paper dates back to 2009 and is titled "Group recommendation: Semantics and efficiency." That work was about defining semantics for group recommendations and how to reconcile different users' perspectives, and how to do that efficiently in a dynamic fashion. That was work that I did with Senjuti Basu Roy, which in fact started a long collaboration with whom I still work.

That work really showed how data management solutions, materialization, indices, etc, can be used to design faster recommendation algorithms for individuals and for groups of people. I had great plans for that work: to serve as a basis for rethinking database architecture, models, and algorithms to handle groups, teams, and communities as first-class citizens. I thought these databases could serve as a backbone for building Web applications. When I joined CNRS, I recruited several colleagues in Grenoble to build SOCLE, a framework for data preparation in social applications, where we used several of those ideas that we had initially. I gave multiple tutorials at WWW, SIGMOD, and VLDB, wrote surveys, and collaborated with experts to add a visualization layer. We also had several accepted demonstrations in the viz and database communities. Despite that, I still feel highly unsatisfied because I did not bring it together into a single system. So many applications and user needs to reconcile, so many content retrieval and recommendation algorithms to bring together, and no one system to rule them all. I went through my paper titles, and my longest recurring words are "group/community" from 2007 to 2023! So

[1] Sihem Amer-Yahia and Mounia Lalmas: XML Search: Languages, INEX and Scoring. ACM SIGMOD Record, v 35(4): 16-23, 2006.

[2] Sihem Amer-Yahia, Ricardo Baeza-Yates, Mariano Consens, Mounia Lalmas: XML Retrieval: Integrated IR-DB Challenges and Solutions. SIGIR Tutorial, 2007.

[3] Sihem Amer-Yahia, Ricardo Baeza-Yates, Mariano P. Consens, Mounia Lalmas: XML Retrieval: DB/IR in theory, Web in practice. Proceedings of the VLDB, 1437-1438, 2007.

maybe I should not despair, and this may happen someday.

I believed we as a community would start rehauling DB systems to handle individuals and groups as first-class citizens, but that did not happen. We are proud, as a community, of building generic databases, and I believe we are a bit resentful (including myself) of building special-purpose databases. We need to talk about that. We need to talk more about our failures. Maybe another Failed Aspirations in Database Systems workshop would be great. I enjoyed FADS@VLDB 2017 very much, and I think I was not the only one who enjoyed it.

*You have done a lot of work, besides your technical work, in terms of contributions to the community. You have initiated the Diversity, Equity, and Inclusion (DEI) initiative in the database community and chaired the DBDNI group for three years. Your DEI work has had a great impact. Were there particular events that motivated you to go down this path?*

I just did not want to attend another women's lunch! In fact, Juliana Freire, the SIGMOD Executive chair at the time, asked me if I could have a DEI working group for SIGMOD. I told her I'll think about it… and then Jeffrey Ullmann received the 2020 Turing Award with Alfred V. Aho. A big controversy broke out on whether or not it was a good idea to celebrate Ullmann and his work as a community. At that time, I was the DEI chair for VLDB 2020, and I tried to put together a panel to discuss that, but I failed. Most people I contacted pushed back and expressed concern about being stigmatized when talking about that.

I felt we needed to take a step back and understand how to approach that kind of question and be less emotional about it. So, I went back to the social sciences and I discovered Gisèle Sapiro, a CNRS sociologist and historian who wrote a book titled "Can we separate the work from the author?". I felt relieved to find a scientist who could give us perspective. Gisèle asked me for all the material I could give her, and she researched Ullman's case and drew parallels with other scientists' cases. For her, that was like devising an algorithm for us. She told us that while ethics is a growing concern in scientific communities like ours, we are not the first ones to ask ourselves the question of the relationship between an author's ethics and their work. Since the feminist and civil rights movements, increasing attention has been paid to sexual harassment and discrimination in academia. Some scientists argue that authors who engage in unethical behavior should be cancelled or at least not rewarded for their work. In contrast, others contended that the work should be dissociated from its author. She outlined a plan on how to think about those questions. It helped to see that one could approach delicate questions constructively, instead of becoming emotional about them. Of course, that is only one aspect of DEI, and we understood that it is both important and fascinating. Today, the DEI initiative is about so much more, and I am glad it has evolved.

***We are proud, as a community, of building generic databases, and I believe we are a bit resentful (including myself) of building special-purpose databases. We need to talk about that. We need to talk more about our failures.***

*Speaking of the initiative's evolution, you had unusual success in terms of co-sponsorship from multiple conferences and societies, which rarely happens with anything. Even recently, we had DBCares merging with the DEI initiative. Can you comment on this? How did you manage that?*

I think all the stars were aligning – we had many great people interested in those matters. But let me first say that it is much easier to get things done at the level of individual research communities and then elevate them than at the level of organizations such as the ACM or IEEE. And for that, we are lucky to have the SIGMOD Executive Committee, the VLDB Endowment, the EDBT/ICDT Executive Committee, and TCDE, all of which adhered to the DEI initiative at different moments in time.

I looked into all the efforts that were happening in our conferences and felt there was potential to reduce redundancy, and sort of build a history together that goes beyond conference boundaries. I initially reached out to people in the DB community who have been involved in DEI events in the past with a proposal to run a meeting and pick their brains on the topic. I just did not want to repeat things, so we had a first meeting. Things grew rapidly from there. I also found that the SIGHCI community was at the forefront of DEI questions in 2020, and it served as a great inspiration for structuring our DEI initiative. I started running one-hour meetings every other month. Because it was during the pandemic, people felt excited about discussing topics that gave them a sense of purpose. As discussions unfolded, it became clear that we needed to define specific actions and designate ambassadors who could

advise individual conferences and help build continuity in our efforts.

Merging DBCares with DEI happened naturally when we started talking about the ethics action. Similarly, promoting the use of CLOSET for CoI detection became part of DEI because it is an ethics concern. In fact, SIGMOD and VLDB will cover the costs of hiring a software engineer for one year to develop a tool to help with PC formation and paper assignments based on CLOSET. I recently learned that the SIGSAC Executive Committee established the Committee on Preserving Professional Conduct and Academic Ethics (SIGSAC PROTECT) with the mission of providing a coordinated and timely response to emerging ethical concerns. Today, many other communities, including NeuRIPS and KDD, are reaching out to us to share our experience in setting up the DEI initiative.

*Among the actions of the DEI initiative, there are many components (e.g., Reach out, Include, Organize, Support, Scout, and Coordinate), some of which you discussed. Is this vocabulary something that people are more broadly aware of? What are the most challenging tasks that you and your colleagues have addressed with a long-term impact?*

This vocabulary is something I came up with because, for the first year, every time I thought about it before our next meeting, I would be doing things. For instance, *Scout* came up because I was really scouting for DEI events, trying to understand what other communities were doing, what was happening on universities' websites, etc. They were evolving and talking more about DEI. I said, "OK, maybe I should just coordinate the initiative and let my colleagues who were part of the initiative to scout, support, organize, include, and reach out". So that is how this vocabulary came up.

***… involving men in the DEI initiative and, more generally, in DEI events, was and still is the biggest challenge.***

We faced multiple challenges. In the reach out action, the idea was to design a single questionnaire and deploy it to every one of our conferences to understand the profiles of people attending our conferences and how attendance evolved over time. Of course, deploying those questionnaires came with the challenge of ensuring privacy – where would we store the data we gather about people? And before that, there were questions about how to design the questions. How do you ask about gender and sexual orientation? Would people be willing to provide that data?

We also encountered issues related to funding DEI efforts. One thing we realized is that we need to plan upfront and make sure conference organization proposals include a line for DEI events in their budgets, so that they are treated as a first-class concern in our events.

Thinking about what DEI means for journals and workshops is still ongoing. These and other challenges are discussed in our yearly SIGMOD Record reports. And while promoting DEI is honorable, enforcing DEI is not always easy – it is also risky, and we are not trained for that. Depending on what we are talking about, a harassment case or a CoI violation case, approaching it constructively without building stigma around the individuals involved is a hard question.

I risk falling into a cliché, but involving men in the DEI initiative and, more generally, in DEI events, was and still is the biggest challenge. DEI is a nurturing and caring activity; men and women have different ways of caring. We need to include both ways of caring in our DEI actions, and to do so, the initiative must include more men. In my career, several male colleagues have had a supportive role, including yourself, Jag. Thank you for that. We need our male colleagues to engage more in generous and empathetic behavior.

*Beyond your own DEI efforts, what advice would you give women entering the field of DB research today?*

First, I want more women. Please go into it! If you would like to do database work, please do! It is really fun, and the people are very nice. I love my colleagues!

I would like to give two pieces of advice to both women and men. The most important advice is to be aware of the fact that mentalities have changed. If you see something or experience discomfort, unease, or shock, you can talk to colleagues involved in the DEI initiative. You should not feel that it is your fault. If you still think it is your fault, it means that the initiative still has a long way to go. So, the initiative is there for you. The other advice is to realize that change does not happen by itself and that everyone is welcome to get involved in DEI efforts.

And to both men and women, an important thing is to have an activity outside of work to let off steam and explore other sides of your potential. Doing research is very personal and is highly rewarding when we succeed, and hard on us when we do not. Defining one's achievements solely through research is not a good idea. One needs to devise ways to compensate for failures. I do it by dancing and keeping in touch with friends around the world.

*That is a good point to move on to more personal matters. You have worked in industry, academia, and research institutes. You have lived and worked in different countries. Do you have thoughts to share on how all these compare?*

For a long time, research in academia and industry was conducted in very different ways. One thing that is quite unique in the industry, particularly in the Web industry, is that people with very different career paths and research areas are striving to achieve the same goal. That makes working together with other disciplines more natural. In academia, historically, boundaries between disciplines have been quite rigid. When I arrived at CNRS, it was the first time I started working in academia. Before that, I had been trained to work with people in other disciplines, and it took me a while to do that again since I joined CNRS.

Luckily, the advent of data science has been changing that. All the AI institutes are gathering people from different disciplines. Most researchers in research institutes such as CNRS and INRIA do not teach, as it is not required. That leaves plenty of time for them to chase funding for their research. However, they have less access to students. In France, we have mixed research units that co-locate University professors with CNRS and INRIA researchers. The lab I work in, Laboratoire d'Informatique de Grenoble, is a mixed research unit where we benefit from each other's perks. I spend more time raising funds for our research, and my colleagues spend more time convincing students to join us.

As for working in different countries, while chasing a paper deadline and other mundane activities we do as researchers feel the same everywhere, I must admit that the experience and sense of purpose change a great deal between places. In my opinion, the industry is rougher. When I was at Yahoo!, I worked very hard and was very excited about the research I was doing because I had access to great data and some of the brightest colleagues, but I never felt I belonged. I had tough and misogynistic bosses. I am not sure they or I were fully aware of that. Luckily, I was very excited about my research. Also, attending conferences and seeing my friends and colleagues was a great consolation, and living in NYC allowed me to make great friends and practice my dancing. That's a big part of my life. I hope things are different today in the industry. I can't really tell.

When I joined QCRI, in Qatar, Ahmed Elmagarmid suggested we put together a mentorship program for undergraduate students to come and spend time in the lab, participate in research projects, and get ready to apply for grad school outside of Qatar, since there were no graduate programs there. I was very surprised to see so many women sign up for that program. It turns out there are many more women than men who study Math and Computer Science in the Middle East and North Africa region. I come from North Africa, and I did not even know that. Most of the interns we had ended up doing a PhD in prestigious universities in the UK and the US. That really gives you a great sense of purpose.

*Talking about the sense of purpose, what have you found to be the most rewarding in your work life?*

The most rewarding thing is learning from other research communities, both in Computer Science and in other fields. Lately, I started collaborating with Education scientists, and I am discovering different theories on how people learn alone and with others. Some of those theories are making it into my recent research. It's amazing to have such freedom. And, probably the most important thing is to meet people from all over the world who think differently, are smart, hardworking, and ambitious, and among them, beautiful people like Divesh Srivastava and Tova Milo, who lift you up.

*You have enjoyed dancing all your life. Can you tell us something about that?*

I grew up in Algeria, and I was around 4 years old when I took my first dance class. Later, I became a member of the Algerian National Ballet. There was a point in my life, when I was in high school, where I was given a choice of dancing more and doing less math, or continuing to do what I was doing. It was a tough decision. I ended up dancing less, and I found locations to dance less professionally.

To me, classical ballet is like Boolean queries in relational databases. Let me attempt to do this parallel. Classical ballet is very well defined. There is this one movement, you have to do it the way it is dictated: your legs are either plié or tendu. The former is a basic bending of the knees while keeping the heels on the ground. In the latter, the legs are fully stretched. So your legs are either bent or stretched. When I arrived in France as a student, I started Modern Jazz. In Jazz, a tendu is a fluid movement that travels through checkpoints without stopping. Your legs are never totally tendu or totally plié. They're always in between. It's more like IR. Everything is a potential answer to a search with a score. When I moved to NYC, I learned to dance Simonson's Jazz, an organic approach to movement that prepares the body to dance in a way that complies with your anatomy. In Grenoble, I've been dancing Horton Jazz, which focuses on stretching in opposite directions and smoothly connecting flat backs and lateral stretches, tilt lines, and lunges.

I feel like my work life has gone through that kind of evolution. In fact, my whole life has gone through that evolution. Living and working in different places taught me to better understand who I am and what I seek, build smooth transitions, recognize what I like in a place and be grateful for it, and approach my life and choices holistically.

***While specialization in science has contributed to remarkable progress, the separation between fields that aim to maintain their distinctiveness constitutes an obstacle to innovation and collaborative efforts.***

*That is an amazing parallel! To close up our conversation, let's go back to technical stuff. How do you see the future of data management research, and what is the next pressing challenge for us as a community?*

We are the data experts, and we know how to deal with data. We have growing amounts of data, including data about people, and that should really help us to understand how to care more about people. To do that, we need to take a step back, understand what "Transdisciplinarity" means, and focus on integrating intellectual frameworks that transcend individual disciplinary viewpoints. Conceptual frameworks from different fields can provide a broader perspective in both research and practice. For instance, in Positive Psychology, there are several theories that can be applied to the field of AI & Well Being, and that have so much to teach us in terms of paying more attention to people when building human-facing and human-caring DB systems. The work on fairness is going in that direction, and I think we can do more by attempting to answer other fundamental questions, such as "How do we capture experience, satisfaction, and frustration that users experience when interacting with data? How to devise data processing algorithms that optimize for positive feelings? The good news is that there are many theories, such as the Flow Theory in Psychology and the Self-determination Theory, that can help us.

From an intellectual standpoint, advancing research in one's field can be significantly influenced by other disciplines' theories, concepts, and methodologies. While specialization in science has contributed to remarkable progress, the separation between fields that aim to maintain their distinctiveness constitutes an obstacle to innovation and collaborative efforts. Practically speaking, the challenges currently confronting our world do not align neatly with academic disciplines; instead, they are increasingly complex, chaotic, and interrelated, and humans are in the middle of that. Consequently, there is a growing recognition of the necessity for a more comprehensive and integrated approach to understanding these multifaceted issues. That can only be achieved by transcending disciplinary boundaries. This situation further supports the argument for reforming educational practices and advocating for a more cohesive and integrated curriculum in our universities.

*Thank you, Sihem, that is a wonderful place to end.*

Thank you, Jag, once more!