Podcast #8

Mark Surman
Markus Lampinen

ML: In today's episode, we talk with Mark Surman, the Executive Director of the Mozilla Foundation who is also a very strong open-source and free internet advocate for the last several decades. We get to touch on themes such as the evolution of the Internet, where we are today, as well as the ongoing trends around Web 3.0, decentralization, distribution, data sovereignty, personal data control and ownership - all of these fun things. Without further ado, let's get into it.

ML: I am here with Mark Surman from Mozilla. Mark, first of all, thank you so much for being here - it's an honor to have you here.

MS: Thank you, Markus, and I am excited to be here and talk about Mozilla, data, and all the things for the developer audience here on the podcast.

ML: Awesome. Actually, why don't we start a little bit on a personal note? I know that you have been working at Mozilla Foundation and the larger community for a while now. Why don't you take us back a little bit on what the journey has been so far and where you are today? What are some of the topical things you get to work on? Maybe one or two things you are excited about.

MS: Way to go! 12 years of working at Mozilla. It's funny that it is my dream job, but it is also a job that has to get done. I remember how I worked in open source, tech for good and social enterprises around the Internet stuff for a while; around the mid-90-ies I ran a non-profit that helped Greenpeace, Amnesty, and all those people online. Back in the early times, like 1994, I was helping people who do good in the world to use the Internet to do that. So I did that in many different ways for 15 years. Then, I saw a job posting from Mozilla that they are looking for an Executive Director to run the Foundation. I thought, "Oh my God, I was made for this thing!" People there were my heroes, and I was using the early Mozilla browser. What Mozilla stands for - that the Internet should be open and accessible to all - was at the heart of everything I believed, and the fact that they were making a browser (it was 2008 when Firefox had become successful). These were the signs that showed that we could do something different. I wasn't there as a founder; I came later. I went to California; I talked to Mitchell (one of the co-founders) after what it feels like to have been talking to 300 people because my job interview was on stage at their audience. I got the job.

[3:20] Really, the story since then (ca. 2009) has been going from open-source and the idea of an open web stack - the things we were standing for and evangelizing, trying to make that happen in the world - to the Internet, that frankly felt good and people felt optimistic 12 years ago. Going to a place of what makes things better, developers were excited; back then, Facebook was relatively new, Google was not even ten years old, and iPhone barely existed when I first started my job. I remember going to a Mozilla event to get an Android phone to get a first look at it. It was a very different time. We were in an era of optimism around technology.

[4:30] The early years were good in that way. My job was to figure out what the Foundation would do beyond what we did in Firefox. We spent a lot of time evangelizing open source in the science and journalism communities, bringing a lot of these techniques from the open-source into other parts of the world.

Really, by 2015-2016, where my attention was, and where Mozilla's attention has gone, has been really different because we have gone from optimism about the web to where the Internet is way more about people lives, and one where we see that as much as the Internet brings us joy, it also is a place we have bigger monopolies and bigger power centers/ in a few companies than we have ever had in any sector of the economy (even more than we had in the oil industry a hundred years ago). So the economic consolidation and effect on things like elections and anti-vaccine with all the side effects… There is a side of the Internet that is as rosy as when I got to come here and work for my hero.

[6:05] S**o a lot of what we think about today is how can we do more than a browser and more with the browser? What could we do with different technologies and work with developers and the public to make the Internet better again and bring us joy? That is thinking about everything: e.g., from working differently with advertising to looking into how we are doing data-driven computing being something that we need to reinvent so that the users are in control and also that more developers and more startups can get in and compete with big guys?** I think that, in some ways, we are now in a much messier world than Mozilla was ten years ago when I started. But that is also exciting. There are a lot of new problems to solve. We know how to take on big scary problems.

[7:05] ML: It is very dynamic in that sense that the Internet is growing to function and to be more popular. That does not mean that the problems are trivial; the problems are real. We need to solve them. When I told our community that we'd be speaking, I also asked them what they would you like me to ask. One of the questions I got was "the Internet has changed many times. Mark certainly has seen it change many times over his career." In part, what I see in this narrative is that we almost accept it; e.g., where we are. If you drop down on earth as an alien and look at data, you start thinking that data is practically run and controlled by only a few companies. Then you start thinking about that and that hardly seems sustainable. Let's call it the "Big decentralization movement" - there is this giant push, optimism, and even excitement around decentralization that has surfaced in the past couple of years. How do you guys view that as a whole in terms of playing into the evolution of the Internet?

[8:23] MS: the decentralization piece is a part of something bigger - it's a kind of a swing-back to a set of people who are not just accepting that we are stuck where we are. These things go in waves, and I would that at least from where I sit, and we are thinking about these topics all the time, there has been a turn in the past few years from the "wow, we're so stuck" to "wow, this is complex" to "oh, there are things we can do to knock down the complexities and change things again." I feel that it is not optimism but an ambition for something different or an ambition for an optimistic future.

There is a great newsletter called Reboot.[1] If listeners do not know, you should. There are a lot of young people who are in tech in some way, saying they don't want to sit where they are. There is one essay by one of the founders giving the genesis of Reboot. I started thinking that I wanted to fight for all those social justice causes, and then I realized that tech has a lot of power. The way how tech is using that power is not the way I want that power to use. Then she and many other people come back to this and say that we should look at the power we have and see how we can shape the world through technology and come at it with a different set of ethics and different set of principles, and different way of working.

[10:15] I see the decentralization as one piece of that: Building into the architecture of the system that no one party can control things. That goes back to the core principle that made the Internet different from television or what made the Internet not something that governments control. We'll see whether or not decentralization can bring that part of it back: there're so many different layers, e.g., governments shutting down Internet connections. It does not matter how decentralized your file system or your cryptocurrency is, if you can't get to the Internet. It is a complex setting, and people are trying different things.

One of the things I am most interested in is less conversation, is a whole space around responsible data governance and thinking about data from a legal perspective, and not how it gets disseminated. This is a lot about looking at cooperatives, data trusts, and different ownership models that could put me legally in control of all the data that all of the platforms have about me in ways that let me enforce my interoperability rights if I am in a jurisdiction that allows me to enforce my rights around the algorithmic decision-making.

[11:55] ML: let's go a little deeper into that because that is something that we've spent some time noodling on. I've read some of your posts and listened to some of your talks where you explain that there is not only one set of data, but that the data has layers: there is personal data

[1] https://reboothq.substack.com/about

that you have and there are inferences that are derived from that data (usage, interactions). It is nuanced, it is not just white and black. We've often talked about this as four layers of data: there is very clear stuff that is yours and there are all the other things on top.

MS: What are the other three layers? Can you explain?

ML: That is a great question. There is the data that you enter, the data about you. Then there is usage data - the data that you have in various applications. Third, there is derived data based on your use and your profile (that's one tier beyond that second category of "usage data"). And then, fourthly, there is an abstraction of data - what is done with that data on an aggregate layer. That would be a simplification of how we think about data. In reality, we consumers have some control over the first segment, but we do not have any control over the last three layers.

[13:25] **MS**: it's like a city:

- my apartment may be in a house;
- then as an apartment association or as a condo owner I may have some rights in the building;
- then there is the street,
- Then, there is whoever builds next door.

Of course, that is complicated too, but we work in ways that we manage to balance those things, whether it is purely as developers - how do we think about those first, second, and third-level facts, and how do we want people to be able to make choices and enjoy doing things and also protect themselves. It's just new, and we haven't really managed to figure out these things as developers. I was just re-reading a piece of our research that we did a couple of weeks ago on TicToc and the German election. You know, TicToc has put themselves forward as being willing to tackle disinformation. So we monitored them in the journal because they were setting themselves up as a really good actor, and you know what, we wanted to see how things work out as you set yourself as a really good actor. They had labeling of what was official election information so you could tell if this was from the voice of a political party or a candidate. Of course, there are lots of cases where this went wrong. That does not mean that they are bad people or that the developers are not smart or not doing the right thing. We haven't figured out how to do that stuff well yet. That is interesting: how do we push ourselves to make healthy cities? How do we make these complex environments work for people? I think we are still babies in this regard.

[15:30] ML: I come at it from a very personal angle of learning about this. At the last company I was running, we worked with financial companies and financial data. We had these regulations come out in Europe, such as PSD2 and open banking; we had GDPR and a lot of those data portability provisions. Initially, we were very excited about this. I thought that data flowing and the open data market would create a lot of value and more opportunities. But in a couple of years, I started looking at it so that we are bloating the entire Internet because we are taking the same sets of data and doing copy-pastes across every single database. I started wondering how many databases can you keep up to date? How many of those are current and correct? One part is the privacy aspect, and the other part is the accuracy of information: If you want the right type of value that data should be representative. That just ended up as an impossible problem.

One of the things we are looking at is a way of decentralization: if individuals themselves can carry one of those four layers of data - yes, over time we want to tackle all of them - but an interesting aspect is that if you can centralize part of it with an individual (=decentralize). Either way, it was an interesting journey for me personally looking at that data. First, looking at the portability as a very exciting piece; and then looking at it and understanding that it is complicated: we are in this situation where even though the intentions are great, the actual nuances of how to deliver it is not that simple. Especially now, looking at some of the things that are going on in Europe, there's a huge push to introduce more of those new business models around data that is individually controlled. That is going to be very interesting to watch in the next 5-10 years.

[17:37] MS: there are a couple of things that make me think about it. It is a fascinating space. One of these is that I am loosely involved in the Data Collaboration Alliance which takes those two concepts and puts them together. It illustrates how tricky this is, but also the new thinking that is emerging. One, is the idea of the *zero-copy data* and, metaphorically speaking, I own all my data. Of course, when we get to the level of inference, that becomes a trickier thing. But that's both - the accuracy and ownership - if we get to a point where it has been thought about within MyData and more simple types where I literally have my own data store - that is not the right architecture to build - but imagining architectures that effectively are with a trusted party where I own or have that relationship with them, you have some choices and the ability to exercise my data rights more easily. It's really exciting to see people poke at that even though it is still very early days.

Developer / open-source community

[19:01] ML: that certainly is. Could you talk a little bit about the community that you guys have and the initiatives that you have going on? Could we think about it less from the technical project perspective, but more from a community as a movement. When you talk about data privacy and so on, what are some of the things that the community cares very deeply about and are working towards that?

[19:30] MS: There are mainly two examples that I'll give you. One, common voice community that we are involved in, that is about creating an open-source data set that is used to train text-to-speech and speech-to-text models ("voice data set"). The other is something called "Data Futures Lab" where people are trying to implement at a practical level some of these alternative data governance approaches. Both are the communities around those themes.

And the Common voice one is so exciting to watch because it is not like a voice that voice technology is difficult to access for a developer - you can get APIs from Amazon, Google, from whomever. But there are things that big companies do not pay attention to. In a way, it feels a lot like really open source communities who at around this voice data set (an audio clip tied to a piece of metadata that says what the audio clip has inside of it) and then the validation process is a pretty rudimentary process. Who shows up for this is so interesting because we have communities in things like Esperanto, or Basque, or other small language communities where they do not see themselves represented in the main technology, as well as really big mainstream languages like Swahili - there is a large community growing around that (tens and tens of millions of people).

So you build a community where people scratch their own itch and put it into something that is bigger than themselves. This community really exists in that common voice setting. I haven't seen that many places yet where the real open source community ethos has shown that open data or AI training data stays. I think that there is a real potential for that because so many things that we get off the shelf (language models, APIs, speech AI, whatever) are designed from purpose-centered assumptions. Whether you care about purpose-centered assumptions from the business model perspective or not, don't meet the needs of the long tail.

We started that project just as a wonky R&D project a few years ago and now it's grown into a huge data set: dozens of languages and a partnership with Nvidia and people in places who try to use different languages. That is one of the places where I see an open source vibe emerging around data and AI that is really exciting.

[23:10] The other is that Data Futures Lab, where people are asking some of these developers some important questions, e.g., how do you take GDPR and think about it not just as a thing that you need to comply with but also how to use it creatively to do different things, perhaps even create companies that might have a strategic advantage that ethical/responsible data regime that is different from the standard practices at big companies. For example, some people at the Data Futures Lab both - a union called "Workers' Information Exchanges" (basicalle, a group of Uber drivers) and then there are consumer reports around privacy - both are trying to figure out how to give people some value through subject access request (i.e., user's request to give data under the GDPR or CCPA), how do they create value and build services that scale by people aggregating data through subject access request from various platforms to create AI and Machine Learning models to understand how to look into the back of the platforms.

That is super interesting, but it's not straightforward: If you think you could have 5,000 Uber drivers, we could build a dashboard or we can see how the pricing works, or whatever. But it is hard to get the data; it's more of a legal process than a technical process. The punchline of this story is that they were both trying to do something in the space, but a bunch of the people on their teams are developers and really practical. They realized that they really need to start working on specifications that eventually can become the standard on how these subject access requests are formatted. Until we get to the point where you can require companies to have an API for subject data access request, you can't actually do anything with this [data]. So have these developers now who are rolling up their sleeves, building those specifications, trying to find friendly companies who are trying out those specifications. It's another example of those cool open-source group around data.

[26:05] ML: there are so many things to unpack here and all of your points are super fascinating. With regard to the notion of data request, there is almost like this contrast where a lot of those companies have public APIs, but those are for their commercial products, and they will only expose a certain set [of data]. Then, at the same time, you, as an individual, can go around that and ask them to give everything. But then, you get a CSV file or a folder, and then you wonder what to do: these are cool tables, but now what? The cost of that entire thing - I remember a report from 2020 - noting that it costs US$1400 for companies on average to comply with one such type of a request. That seems like a perfect storm that will mature over time.

You are absolutely right that it feels that there is not enough of a pain that there is an immediate need for a company to figure out what they need to structure to make this happen. How do you see that playing out?

[27:08] MS: Well, you have to look at all of the things together: what consumers want, how developers think, how business and product managers think. But if you step into a platonic lab environment and think about the things that we are talking about. Why does it cost US$1400? Partly, us too, we have problems with figuring out subject access requests. We don't collect very much data - that's part of Mozilla's motto - but we still collect some data, and people do send subject access requests. It is sitting in five databases, it's not sitting in a single format. From our perspective, if we are giving them something, it has to be good. Why? Because we think about people, and data structures, and iterative approaches that other companies have never thought about people's data rights, or people's expectations of interoperability, or people's expectations of access.

If you go back to file formats before certain standards emerged in the market (such as rtf), it was the same thing: you never thought that “Oh, I would design for portability” or you were doing was trying to lock people in, or simply because it was not important to give people a product. I think that one of the things that will hopefully happen both out of consumer and regulatory pressure, but also over time as developers and product managers start seeing utility, it is more design for the idea than in the end people are the objects in our data spaces, that if they are being able to see it then we can build value around that. SO I think that it is an interplay between the practices of developers and the way we think about the architecture of these things from the get-go, and the way how we pay our technical debt and think about refactoring things over the life of a product, and what do the consumers want and regulators require.

[29:40] ML:

[... more coming...]