Tim Phillips [00:00:00]:
Today on VoxTalks Economics, an ambitious new project to dig up microdata and make it accessible to researchers. Welcome to another VoxTalk, recorded live at the Center for Economic Policy Research, Paris Symposium 2023. My name is Tim Phillips. Microdata that's collected at the level of a household, a firm, or even an individual is extraordinarily valuable for researchers who are interested in how policy works. But although it sometimes seems we're drowning in data, there's not enough affordable, accessible, high quality data of this type to satisfy researchers and policymakers. Last year, when VoxTalks visited Paris, we spoke to Filippo di Mauro of Compnet and Ugo Panizza of the Graduate Institute in Geneva, in a hotel bar, I seem to recall, about their ambitions to improve this situation. So I'm catching up with them once more to find out what progress has been made. Welcome back, Filippo.

Filippo di Mauro [00:01:16]:
Thank you.

Tim Phillips [00:01:17]:
And Ugo as well.

Ugo Panizza [00:01:18]:
Thanks Tim.

Tim Phillips [00:01:20]:
You call microdata a public good. So can you give me an example of why this is so, a policy question in 2024 for which micro data will be important?

Ugo Panizza [00:01:32]:
Okay, Tim, so let me maybe use next generation EU. So that's a massive program. The European Union is planned to invest more than 800 billion in this. There are several objectives. Make it green, make it digital, make it healthy, make it strong, make it equal. Let me pick one of them. Make it strong. So the pillar of making strong is to encourage young people to study science and technology, which opens door to green jobs in the future, supporting education and offering loans and grants to young entrepreneurs. That's great, but how do we know what works, what doesn't work, how do we know how to do this in practice? So this is where microdata can help. Having detailed information on firms, having detailed information on people, having detailed information on education will help us evaluating what are the most efficient way to reach these objectives. And to do this, we need microdata.

Filippo di Mauro [00:02:24]:
Let me add one thing on competitiveness is the catch word of 2024. Apparently the commission president has established that as depriority. Mario Draghi has been tasked with a report due in June. How is it going to be done with the lack of data that we have at a micro level? I mean, imagine that you want to check how the enterprises were affected by the COVID and by the say Ukrainian world and understanding how differentiated it was, the impact between large firm and

most productive or firms that were involved in global value chain or not. Well, you're not able to do it. Maybe you can do it for France because they have fantastic data connecting productivity and results of the firm with custom data, but you hardly be able to do it with other countries. So good luck to Mario Draghi that unfortunately.

Tim Phillips [00:03:18]:
Almost everything we do seems to create data somewhere. So is our problem here that the data is just not available, or is it that it doesn't exist?

Ugo Panizza [00:03:30]:
In most cases, it's just that it's not available. Of course we can think for some problem which we don't have data, but for almost every question I can think of, I think there are data around. The problem is how do we access this data, how cost is to access this data? Maybe there is some data I would like to have that don't exist. But if I could access what exists, I think I could answer 99% of the questions that I have in my head.

Tim Phillips [00:03:52]:
So give me an example what micro data could potentially be used that exists at the moment, but it's hard to access.

Ugo Panizza [00:04:00]:
Let me give you an example of a country, and this is not even a rich country, it's a middle income country, which has the most amazing data researcher might think of. So this is a country which has data on every single firm in the country, and it's linked with every single bank in the country. And we know the loans from one bank to a given firm, and plus the firms are linked to each other through the VAT registry, and plus the banks are linked to international banks through international loans, and plus the firms are linked to the workers through the payroll. So imagine this data set. You can follow a worker through a firm, through a bank, through other firms, you can study incredible things. And this data exists already. The problem that they're very hard to access this country, as, let's say, a complicated political situation. I started a project, and then the manager of the institution that controlled this data changed and the project was stopped. And so, first of all, this model could be replicated by more countries. And plus there is the issue of guaranteeing access to researchers to this type of data.

Filippo di Mauro [00:05:01]:
Let me add on that. True, the data are there, and in fact, we have good news in certain countries, we have France and the Netherlands in particular, where data access is entirely transparent, is actually subject to fees to remunerate, of course, the statistician that do it in job and access is guaranteed for the researchers that have a good project to follow up. So that's great. The problem is that this is not generalized at all in Europe. And in particular, there is an aspect of legal interpretation of the data protection, which is basically extremely heterogeneous across countries. And this is really unacceptable. If you think about that Europe is a common market. So this is something that we need to correct. There are countries in which the data are

extremely good, but basically have data protection, access to the data, which is subpar with respect to the other countries. In particular, one mention is that in certain countries, you have access only to national researchers. So there must be a researcher of that same country that has access to the data. This is not acceptable because there are other countries instead in which these procedures are very different and confidentionalized. Every European good researcher can access the data across certain confidentiality conditions being maintained.

Tim Phillips [00:06:22]:
You've been busy since we spoke last year, and one of the things that you've been doing is asking researchers what their problems are in getting access to micro data. What have they been telling you?

Ugo Panizza [00:06:33]:
Well, there are all sorts of problems, and this varies on the field on which individual researcher work. One common problem is cost. Some database are provided on commercial basis and are incredibly expensive. And this, among other problems, creates inequality in research. Right? So if you belong to rich institutions and you have a big budget, you can do certain things. If you belong to a smaller, poor institution, you don't. So cost is one problem. Some data sets are free to access, but certain people have access. So Filippo mentioned that if you're a citizen of a certain country, you have access to certain data. That's a problem. Sometimes it's just who you know. If you know the right person, you get access, and if you don't know the right person, you don't get access. So there are several problems, and each of them needs to be dealt in its own way.

Filippo di Mauro [00:07:28]:
Which is actually, again, the point of Ugo is that you have this data sparsely available. And therefore, when you really want to do some cross country study in Europe, you have to use very costly commercial data, which, by the way, and I'm sorry to say, are not necessarily the best quality because these are data that are available, publicly available. But of course, since we said that there is a problem of access to the national statistical institutes obviously are not the best because the national statistical institute in certain countries would have much better data. So our aim should be to basically get out this data which are in the vaults of the national statistical institutes, therefore lowering the cost for everybody, rather than getting the shortcut, which is, okay, let me use the data that is available. I'm rich researcher, so I can afford that. That is not good.

Tim Phillips [00:08:19]:
Now, there are many organizations in the last few years that have made giving access to their data, the data that they collect, just as a byproduct of whatever they do day to day, part of their business strategy, monetizing that data. If we say that this data is a public good, should access to that data for research be free or be capped at some level?

Ugo Panizza [00:08:42]:

Public institutions, I think they should have a duty to provide this data either for free or without making any profit at cost, because that's part of their mandate to produce public goods. Now, if these are private for profit institutions, it's hard to tell them what to do with something which is their property. But I think it would be ideal and probably also convenient for them to have clear pricing strategy for researchers, because now some of these institutions deal with big pop up private sector firms and they say to access this data, it costs half a million dollars. If you're a big bank, it's okay. If you're a researcher, it's not okay, and then you have to call them and each person gets a different price. So if this institution were to have clear academic pricing or agreement with consortium would be definitely a good step.

Filippo di Mauro [00:09:29]:
I think, Tim, financing is a key issue. Statistical institutes are structurally understaffed, underestimated, almost considered not at the par of the economists. Right. So this is something must be changed, because the average quality of their work is just amazing. The richness of the data that they have and the quality of the stuff that they have, although they are entirely underestimated, is fantastic. I have incredible experience in several countries. So for me, the financing issue is key, because we have to make sure that they are no longer understaffed chronically that are able to provide the services when it's needed. I mentioned the case of France and the Netherlands, is a great case, because actually they have been monetized. Their services, they're not terribly expensive with respect to commercial data sets, is just a fraction of that price. So I really would go for a systematic way of financing, and therefore a participation of the researchers in that. Using the funds that are normally available for data set funding, etcetera, I think would be absolutely the right way to go.

Tim Phillips [00:10:30]:
Of course, in the private sector, you can't make people do anything with their data. In the public sector, though, there is the opportunity to have a standard process for access, standard costs.

Filippo di Mauro [00:10:40]:
Exactly. We want to have this transparent as it is in the France and Netherlands. Again, this is a template that Europe should follow.

Ugo Panizza [00:10:49]:
So, for instance, the Italian Social Security Administration in 2015 created the Visit INPS program. Not everybody can have access to the data because they have limited capacity. But there is a clear process. I go on their website, there is an application procedure. So everybody has the same information. So that should be the gold standard everywhere, to have clear procedure everybody's aware of. And then it depends on the institution, of course, and the type of data. But the fact that everybody should have the same type of information and in principle, the same ability to access the data, I think is very important.

Tim Phillips [00:11:27]:
When we spoke last year, you had the idea to create like a micro data hub for Europe. How are

those plans progressing?

Filippo di Mauro [00:11:36]:
Well, Tim, fortunately, they are progressing very well. I must say that we have been getting a financing by the European Commission, DG Reform finance, a so called technical support intervention mechanism. And this provided resources to hire a few staff, young economists that help us out to be an intermediary between the statistical institutes, the national productivity boards, which will supposed to help train and have assistance to use appropriately this data set. So we are making tremendous progress. We have already five countries which have set up the microdata infrastructure, the MDI, and we expect few countries more to come up. Yesterday, I just spoken with the ISTAT, the Italian ISTAT. They have been guaranteed that in a few days they're going to be on board. I was meeting in Malta. They are going to be on board in a few weeks. And we have Germany already, and France and the Netherlands, as mentioned before, are already the champions for us. So we are really proceeding well. Thank you very much, by the way, to the CEPR community for helping us out on that endeavor.

Tim Phillips [00:12:38]:
I can imagine this is not an easy problem to solve. How does the MDI work?

Filippo di Mauro [00:12:42]:
Yeah, it's not easy at all. We establish access to the data country by country. There is a large differentiation of access procedures. So what we do, we keep up having bilateral contacts with national statistical institutes, making sure that we sign up a memorandum of understanding on the kind of data that we want, the kind of procedure that we have to follow to have this data. And then at that point, we create codes which are basically harmonizing the data collection, so that when we run a code for, say, France, it's the same code for the Netherlands. In principle, what we would have is a situation in the steady state whereby a researcher will come to us with a set of codes, say, on the productivity export dimension, say, and we will deliver output out of that code in a few weeks for the number of counties which are participating.

Tim Phillips [00:13:36]:
So what's the end game here, Filippo? Is the MDI designed to be an open data store for everybody, or is it for specific researchers and research projects?

Filippo di Mauro [00:13:46]:
Well, we have to distinguish two things, right? So we have on one hand the component data set, you may recall. Okay, I mentioned to the public what it is. It's basically a set of indicators that we build every year which are set, and there are like 300 or 400 indicators which study the relation between productivity and the drivers of productivity. So this is a fixed data set. That one will be public. Now, the MDI instead, will be a much more modular kind of tools, because at that point, although we have an underlying data set, it will depend on the kind of questions that the user will post to us, right? So in that case, it will be much more modular and much more personalized. So every two, three months, we will run a set of codes, but this will be different

from the ones who will be running just a few months earlier. So that's the concept of the difference on that.

Tim Phillips [00:14:35]:
Every day in the news we're hearing now about AI and big data, and it's made people much more aware of how much data there is, how much data there is on them. How do you think that this is going to affect your project? Do you think it's going to make people wary because they are conscious of the amount of data that's captured and shared about them? Or do you think it will make people enthusiastic because they realize the potential of that data to help their lives?

Ugo Panizza [00:15:01]:
I think it will be mostly an opportunity. So most of us have read the book by Daron Acemoglu and Simon Johnson on AI, and their point that AI can be used for good and for bad, and so I hope it'll be used for good. And so I think in term of access to data, we can do amazing things, partly with the data that it's already there when we talk to economic historians, their main challenge, so there, the data exists, but they are in some archive, written in some weird language, hard to read, and now economic historians, they are transcribing things. There I can think AI having really be a game changer, because if we can just take picture and have AI recognize this thing and digitize the data, there could be really something incredible we could do.

Filippo di Mauro [00:15:44]:
I totally agree with the potential that Ugo underlines. However, we are not at that stage as yet, because at this point, what we are aiming at is to create the basic data for the most countries as possible. Eventually, however, when the data set, the MDI is ready, of course, the artificial intelligence will help us out to manage the request, for instance, or making sure that we can cluster the request so the system will be tremendously effective and tremendously more efficient because of the AI. So I think that's tremendous advantage for our project.

Tim Phillips [00:16:16]:
So, final thing then, Ugo, what's the CEPR's role in all of this? What can it do?

Ugo Panizza [00:16:22]:
So our president Beatrice, always tells us that C stands for center but should be N for network, right? Because we are really a network of researchers. And I think this network can really play a role in improving data access by doing many things. First, as a center, CEPR can coordinate people, but as a network, we can inform researchers, we can inform people, and we can inform policymakers. And that's really key because as Filippo said, resources are key. And at the end, these resources, public resources, are controlled by policymakers. So if policymakers fully understand the benefits of having better data, for instance, for evaluating next generation EU or evaluating competitiveness, then we'll be willing to put resources in statistical offices and create incentive for this statistical office to allow researchers to have access to data. And I think here, the CEPR network, the CEPR community, which is very visible in individual countries, can play a

big role.

Filippo di Mauro [00:17:20]:
For me, what is very important is also that we try to convince our colleagues to change the mindset. The idea of having a project which is only on one country, possibly small country, is not necessarily the best way of going about. We really need to go as much as we can and as soon as we can into a multilateral data collection mechanism and research based on that. I mean, they do it in the US. They don't do so much of a state, Missouri versus Michigan. Let's stop doing it in Europe. But this is, again, it's not CEPR. But I hope that actually many people share my view and will work on that.

Tim Phillips [00:18:00]:
Huge opportunity, huge challenge as well. So thank you, Filippo.

Filippo di Mauro [00:18:03]:
Thanks.

Tim Phillips [00:18:03]:
Thank you, Ugo.

Ugo Panizza [00:18:04]:
Thank you, Tim.

Tim Phillips [00:18:13]:
If you want to read about the research that's been done so far, then there is discussion paper, Microdata for Economic Research in Europe. Challenges and Proposals. Many authors, Filippo and Ugo are two of them. It is discussion paper 18640 at CEPR.

[Voiceover] [00:18:34]:
This has been a VoxTalk recorded at the CEPR Paris Symposium. When you're checking Instagram, look for clips of past episodes at VoxTalks Economics. Follow us wherever you get your podcasts. We're there. And if you like what you hear, why not leave us a review, too? Next week on VoxTalks Economics, does profit sharing make employees or firms better off?