Interview 20

Interviewee	22-User-A
Interviewer	Ashraf Shaharudin (TU Delft)
Date	10 August 2023

Interviewer
OK, so my first question to you is of course, could you please introduce yourself, your current role and your professional background?

Interviewee
Thank you. I’m <redacted>. I'm a <redacted>. My background has been geomatic, engineering or geospatial analysis in <redacted>. For my <redacted>, I did <redacted>, so immediately after <redacted> or after couple of months I went to [country A], where I did my <redacted>. Then I joined <redacted>. In my current role, I look at your spatial health for different health outcomes. But at the moment I'm more interested with maternal health in urban areas, and using geospatial analysis to get more insights.

Interviewer
And your research is mostly, in which areas, meaning the geographic geographical areas?

Interviewee
My work has been in <redacted> countries and I've worked a lot in <redacted> because I'm a <redacted>. But I've worked in majority of Sub Saharan African countries. I have researched Africa level study or regional studies where maybe it's East Africa or a couple of countries in West Africa. At the moment, I'm working more on <redacted> countries and I think around central, like <redacted>, <redacted>, those sort of countries. So generally Sub-Saharan African countries.

Interviewer
And do you use open data in your work?

Interviewee
My work has been purely driven by open data. I don't remember if I had to pay for any data whether spatial or nonspatial, so majority of the time, I rely on data from third-parties that have been made publicly available, either through publications or maybe an organization has decided to put their data to be open access for people to reuse it.

Interviewer
And so typically this data you obtain from the source itself instead of say from Esri platform? So you actually look for the source?

Interviewee
It's both. It depends whether I can find that dataset from an Esri source or maybe I could find the same data, say it's climate data, maybe I'll go like for <unrecognisable> data or <unrecognisable> or go for <unrecognisable>, but if I think of, say, subnational boundaries, I’ll think maybe of HDX portal. But if I don't find the HDX, I want to see does Esri have this kind of data, like recently they have been hosting like the Sentinel land use data, which is very high resolution. And I use a lot of it for spatial access work, so it's on a use case depending where I’m more conversant with or where I feel that

data might be available. So I use many sources, including the Esri Living Atlas. Actually a lot of the basemap, I rely on them. Maybe I'll use ArcMap and I want to have a base map just to get an understanding of the area. Just to overlay with my vector boundaries, I tend to use a lot of the base maps a lot.

Interviewer
And do you have any challenges or grievances or unmet needs in working with open data?

Interviewee
Sometimes I'm able to like get a layer from them, which I can open from the desktop site. But that data is hosted in a way such that I can display it in my maps, but I can't access the actual vector files, for example.

So there are some data sets that they tend to operate that way. [Some datasets] I can directly get the data straight and use it, but others I'm not able to get the actual layer, but it's like a service so you can display, but you can't like export that vector file, save the boundary data for me to go use it in other software or manipulate. That was actually recently when somebody was interested in some data from Ethiopia. So I have ArcMap and I accessed on their behalf, but we couldn't really actually get those vector files and export them say like to QGIS, where that person is more conversant with.

But when I feel a source is not fulfilling my need, I move on to check the next one because I tend to think we have so many any open source data like if it boundaries for Ethiopia, I can easily go like to geoportal and I'm going to get them there. So if one does not suit, we move on because again, it's open source and nobody is like really responsible for that data because it's free. I did not pay for it, so [they] don't owe [me] anything, so I just have to sort myself and move to the next available data set.

Interviewer
In general, because your geographical scope is mostly sub Saharan Africa, do you think open data from these countries are typically quite hard to get or it's quite accessible?

Interviewee
So I would say it depends on what type of open data. There are data sets that are easier to get, for example in terms of boundaries which I use a lot. If you are looking for national or first or second-level admin boundaries, it's very easy to get those datasets. But if you want to get granular data in terms of like the fifth level of the administrative boundary, you're not going to get that data set.

But if you come to like Europe or the US, you'll find they have, I think they call them zip areas or local areas. You find those shapefiles are well packaged, some people have packages for it. But in sub-Saharan Africa it's unlikely for you to be able to get like at level four and five.

For that data sets like population, so while <unrecognisable> has enabled us to get it easily, also environmental data, our satellite data, it's easier to get those except that you'll only get up to a certain resolution. If you want the high resolution, you have to then pay for it, or if you want more <unrecognisable> and high resolution you have to get money for it. So resources in our context are not that good. You have to work with poor resolutions because of that limitation of accessing some datasets.

Interviewer
And you spoke mostly about boundary data, but since you are also a spatial epidemiologist, I assume that you also use a lot of health data, like you mentioned about maternal health data. So those data are also easily to get or it's quite challenging?

Interviewee
So I use that data of household surveys commonly and those household surveys are done routinely three to five years from Sub Saharan African countries and they are done by the DHS program but also the mixed program. They package those data and they are publicly accessible but you have to register and say why you need those data sets.

So I've used those data sets for like over 10 years and they cater for things like population health, nutrition, social demographic data and all these data are linked to a location so they have GPS coordinates for where they collect those data from, the different households, but they scramble their location to preserve confidentiality and privacy. That has been my source. It's easier to get.

A lot of dataset on health that I tend to use is data from routine health information system. So the HMIS. Those data sets are not easy to get because countries need to give you access to the your data sets because it's packaged in the district health information system, DHIS2, but very few countries will allow you talking to the DHIS2. So for example, I have access to the <redacted> DHIS2, just because of my country, but if you want access to another country, despite having a proposal, it's just research, it’s not profit-making, you won't get access unless you collaborate with somebody within that country, so they probably get access on your behalf. So data from health information system is what is more problematic to get.

But data from household surveys and specific surveys like AIDS Indicator Survey, malaria indicator survey, all those surveys are easier to get. I would know if I request access today, by the end of the day today, I'll have access to that data and I can do all sorts of spatial analysis and spatial epidemiological stuff using those data.

So those are like my to-go data sources. Also census data. There is this organization, I think in the US, called IPUMS, they package all the census data from all over the world and you can access those data sets from them. It's usually percentage of all the census data. So if I want census from Nigeria, I'll definitely go there and easily get that datasets.

Interviewer
For free?

Interviewee
So yeah, for free, no charges. You just need to say why you need those data sets and they confirm the logical use, not for-profit, and it's not misusing the data set.

Interviewer
OK, what's the name of the organization again?

Interviewee
So it's IPUMS.

Interviewer
IPUMS.

Interviewee
So I think it has meaning that I tend to forget from time to time because I use the data a lot. It's something to do with microdata or something, but they provide not only the census data, but they also provide a lot of other demographic and health data that they can collaborate with different organization like they become like the SVR<?> trust.

So for example, SVR trust will host data like Sentinel, not to mean that they produce the Sentinel data, but they just host it to make it more accessible. So IPUMS also hosts a lot of demographic data to make it more accessible and they help you to harmonize with other datasets. So for example in <redacted>, census done 50 years ago and we have our census done today, so they would make sure like the boundaries, you do know which boundary was divided to mean which boundary. They try to harmonize it for you so that it's easier for you to access and use it. So it's like making data more accessible and easier access.

Interviewer
So they are one of the data intermediaries then? Because they're not the original data providers, but they facilitate the data sharing.

Interviewee
Yes, yes, yes indeed.

Interviewer
So my next question is gonna be about Esri as an open data intermediary. My first question to you is how do you use Esri products in your work?

Interviewee
Mainly Esri product in terms of data is I would say, health outcomes or covariates. So for example, if you're modeling travel time to healthcare, you need land use, land cover as one of your covariates to the work. Say if you are modeling malaria prevalence, that becomes a covariate.

So majority of the time when I'm using the data is to act as a covariate or predictor to help outcome that I'm using and that is in addition to the boundaries, because these boundaries are national boundaries, are used for decision making in a lot of countries. So somehow to summarize my results into meaningful subnational areas. So either as a health outcome, as a covariate, or as a boundary to summarize or aggregate my continuous maps. So for example, if we have a raster, summarize it into the sub national boundaries that a policymaker in a country can make sense of. So I think that's how I tend to use the date, mainly for those 3 purposes. Of course, in addition to using their software.

Interviewer
And so you mainly use the ArcGIS product in for data analysis. Do you also use it for visualization or sharing with your peers to use?

Interviewee
I use ArcGIS a lot for my maps. My maps are from ArcGIS, or at least or 95% of the maps that I have made of from 2008, when I started using ArcGIS. All my maps are to communicate for research

publications, for reports, to share with policymakers. All my visualizations are from -- so I use it a lot for that.

Interviewer
Do you also use ArcGIS platform in conjunction with other tools, say Python, or I don't know other tools?

Interviewee
Yes, yes, I do use. I don't like ArcGIS for geostatistical analysis. So I tend to do some analysis in R or STATA, then take that product and go and map it in ArcMap. So ArcMap for me is more of a visualizing tools or data management tool, so I manage a lot of my data in ArcMap, for data preparation and data visualization.

But for analysis, I tend to run away to another software, but lighter analysis like geoprocessing, that is like if I want to do buffering or if I want to do a spatial join, those basic things like a click of a button and it's done in ArcMap, I do it there. But if I want, for example, to do interpolation I will not go use threading in ArcGIS, I'll prefer to go in R. But the data before I take it to R will be a lot of processing in ArcMap, for example, you want to change a corner system, it's just easy to do that instead of me going to the other side. Or I want to merge boundaries or reproject other stuff, will happen in ArcMap then take it to R just for modelling, doing some statistical stuff, then returned back to ArcMap to do the visualization or post processing.

Interviewer
I see. OK. And so in terms of data collection, sometimes you also take a data from Esri or ArcGIS, sometimes you take from outside sources. But then for analysis, you mostly do it outside ArcGIS platform. But for visualization then you use ArcGIS platform.

Interviewee
Yep.

Interviewer
Is there any particular reason why you find it difficult to do data analysis in ArcGIS platform?

Interviewee
So I'm not say it's difficult, but I think the tools are limited. Because what I do outside of ArcMap is a lot of geostatistical analysis, so if I do kriging in ArcGIS, for example, and I use another tool to do kriging, say in R and I use a package, the results will allow me get insights than just putting it in -- yes I can do it in ArcMap and get a result very quickly, but I don't have all the breadth of the results that I would have gotten in R for me to get more insight. So I think it's not made – it’s not tailored to some of the things that I need, which I can tailor myself in another software and that is R mostly because that's where you can like define what you need specifically and how you need it.

Interviewer
OK, so are you familiar with Living Atlas and also ArcGIS Hub? You're familiar with that?

Interviewee
I've seen that name very recently, like two years ago and I started using for especially land use land cover data. But have not explored it more because my case are more of need basis. Do I need a certain link? I’ll find it in my usual source. So it's not like a my first to go place, but I consider it more

for land use, land cover data a lot, but it's not what comes into my mind when I want to access certain dataset.

I think HDX, the OCHA HDX portal, it has a lot of data deposited by so many people, so I think I should get my data set there. If it is not, let me see what Esri are having. So I tried to, also there is for example is OSM or a country, they always give you a baseline data. So I try to look around and if I find it in the Living Atlas, well, good. If not, whereever I find it, it’s what I use it.

Interviewer
So it's not your first go to a data source. And so far, in the limited use that you have probably used data from Living Atlas, do you have any issues with using their data? Do you have anything that you don't like about using their data?

Interviewee
I would not criticise them because if it was my to-go place, I'll have a lot of experience with it, but since I explore as many sources as possible, I can't like can point them for something they have not done because I don't expect them to provide for me everything I need so I don't have like a realistic criticism of the products that I can say for now.

Interviewer
Before we move on to the next question, I just want to clarify, you mentioned HDX, right? What does it stand for?

Interviewee
So it's a very popular data source. So it's humanitarian data exchange. Yeah, it's by OCHA, which is United Nation Office for Coordination of Humanitarian Affairs. They’re not generator of data, but they assemble a lot of data from so many organizations, from so many individuals. It's there for you to use. So they have data to do with all the country boundaries that you need, they have also population data, relative index pollution data, [other] data they find out there. They try to put it there for people to use, so it's another like big source of secondary data sets.

Interviewer
And they are spatial data?

Interviewee
No, not only spatial day, they have spatial data and non-spatial data?

Interviewer
OK, that's good to know. So my next question is, do you have any experience of using QGIS?

Interviewee
I have used QGIS but not as much because ArcMap have what I need. There is no need at all for me to go to QGIS. I tend to go to QGIS when I'm showing somebody and they don't have access to ArcMap and they need to do something. So that's when I only go to QGIS to find that function and orient them to that function, but that is not my daily software. I rarely use it and I have not been forced to use it. I didn't find something in ArcMap that have to force me to go to QGIS. If I have somebody who I'm working with they don't have ArcMap, that's only when we talk about Q GIS.

Interviewer
And so far from your personal experience, peers in your field, from spatial epidemiology are typically very familiar with ArcGIS? So they use ArcGIS a lot compared to QGIS?

Interviewee
It's very different actually. In my field of spatial epidemiology, there are people who use QGIS day to day, and I think it's because it's free. Then there are other people who use purely R for all their work. You can make very nice maps in R, but I don't do that. They are people who whatever they are doing, they will only do enough. So in spatial epidemiology, I know of certain group of people QGIS people. I know another group which is basically R. I know another group is QGIS. Another group is us using ArcMap.

So spatial epidemiology, I would say that based on probably where you are trained and how you started, because if somebody comes to me right now, they like to use QGIS instead of ArcMap because they can't use ArcMap. But where I started in my undergraduate in 2008, we started with ArcMap on our first tier. So I guess I got to know a lot of things about that software. If you go to some universities, I know they don't even have access to ArcMap at all, so it’s QGIS.

So I think it's how they were trained and if they have access because if I didn't have access to ArcMap, I will be using QGIS in my life. So yeah, so it's mixed.

Interviewer
In general, based on your experience, do you think Esri plays a role in enhancing access, supply, or flow of open data?

Interviewee
Yeah, I think they play a role getting open Access data because I've used some of their data and they make it easier to access. So I would say yes.

Interviewer
But you would say that in your case, you also rely on other open data intermediaries?

Interviewee
Yes. I rely on other sources, Esri for me is one of the sources of these dataset and not usually my go-to data sets but if my search did meet with them, I end up using their data.

Interviewer
And do you think Esri also plays a role in connecting other actors in the open data ecosystem, meaning does it facilitate you interacting with other actor with other users of open data or with the data providers? Do you think they have that kind of platform that allow you to do so?

Interviewee
I think the way I see it is they allow me to interact with other data providers because I know in majority of the times they are not the one likely who produced that data set, but they are intermediary of the data set. So sometimes, I want to go to the source of that dataset and maybe see more information, for example, I want to cite a paper that have analyzed that data or derived that data, I know there must be a paper. So they allow me to know more about datasets and producers and if for example I go to the producers of a particular data set that they are intermediary of, I will know more about the dataset that that other person produces. So I think they connect me to other

datasets, probably that I need to know and other actors in industries, mostly researchers that I need to know about.

Interviewer
But do you see any negative or less than ideal impacts of Esri towards other actors in the open data ecosystem?

Interviewee
I think I see it this way. Most of the datasets, you need to have ArcMap to use, and that's a negative thing, because people need to be able to access the data irrespective of the platform they choose to use. So in my case, for my organization this week, somebody needing access to the data set that ArcGIS have, but the first point is you need ArcGIS Online or you need it on the desktop for you to open that data and play around with it.

So yes, they give you the data, but it comes with a with a tag on it, you have to be my follower to use this data. So I think that's the only negative side I see to it, if that data set can only be opened by ArcMap in the first instance or maybe ArcGIS Pro Online, so if that's the first software that you have to use to access that data, I think that's a [negative]. They know they are doing that intentionally.

It's not a good thing because I think there are some datasets but you have to have ArcMap to open, others, not a must. But the one that I particularly had experience with this week, you needed to have ArcMap to open the dataset.

Interviewer
So it's different than your experience of getting data from other intermediaries like HDX or IPUMS where you can get it freely and you can decide what platform you want to work the data with?

Interviewee
Yeah, HDX will give you like five file formats of data sets for the same data. Then you choose whatever you want to do, choose to go to QGIS, choose to work in R, all up to you. Choose even STATA because STATA also has spatial data. Up to you. So they should say OK, you can visualize on ArcMap here, but if you don't want to, you can download these data and go use it wherever you want to use it. If they did that, I mean it would be better and people would use their products more.

Interviewer
Based on your experience, because you were based in the [country A] for a while and you are now in <redacted>, but you are doing research of countries outside <redacted> and <redacted>. Do you think for example, because the way that Esri does is that they have a sort of like a franchise model actually, but not exactly franchise where <redacted>, <redacted>, <redacted>, so individual Esri in each country actually collect data from the country itself and then act as an open data intermediary. Do you think then it doesn't help that for example, you are based in [country A], but you need data from <redacted>? So do you think then in that regard then [Esri distributor in country A] is not responsible for data from <redacted>, so the way that Esri does internationalization, is that it still limits to individual country, do you think that it’s a problem for you?

Interviewee
No, for me it's not a problem because Esri is in Africa. That is what I've been used to, and because I work for Africa, so I have not needed to like work with [Esri distributor in country A] for example. I've

not been forced to because if for example, I come to <redacted>, I don't need to really interact with them a lot. It's only this time that maybe I'm going to interact with them because the desktop that I'm going to have here, I'm going to interact with them and get the license and such things. But even in the [country A], in the university, they have ArcGIS Online, but once I actually have ArcGIS Pro, I can just access whatever I need. I didn't see like whether it's [Esri distributor in country A] or Eastern Africa, for me it was not a barrier, and actually it was not even in the picture that I'm now dealing with a different franchise of Esri. So for me it doesn't play out for the work that I work.

Interviewer
How about in terms of data? Do you see that there are differences? For example, do you see there are a lot more data in the Living Atlas from certain countries compared to, say, developing countries? Do you see that?

Interviewee
I wouldn't say that because data for Africa are more or less similar. If it's not available, it's not available. So I have not had to use data for [country A] to do something, I have not had to use maybe data from the Netherlands to do something, but I know those people collect the data well, they are well organized and structured. But in terms of accessibility, I don't know if Esri will offer you -- how it compares to Sub Saharan Africa or Africa. But for Africa generally, it is known that it’s data poor and data organization is not good. We don't have, like national spatial data infrastructures that are well organized. So I can't compare these regions because I just have experience more with a single region, and our region is more homogeneous. Whatever you get in <redacted> is what you're getting in <redacted> and <redacted>. Maybe only <redacted> is a bit different. They are more developed, they are better organized than you'll find in the other African countries. But sub-Saharan countries are more similar in terms of the data landscape.

Interviewer
My last question. What is your general view about how open data intermediaries, not limited to Esri, can play a better role in the open data ecosystem?

Interviewee
I think the way they can do it better for me it would be very simple, just to make sure that they are using file format compatible with different processing platforms. So the file format in which they present the data set, that's one. Because sometimes you get a single file format which you need to undergo 20 processes to make your data usable in your normal platform.

They are many -- like geotiff, which is a very common file format for rasters, but you'll find climate data is not in that format, it’s only stored in a single format, then they give you a manual of how to convert to geotiff. So why do I need it? If they have the way they can do it and provide us with that. So I think if they can not be very rigid and use the more common formats to provide [data].

Also to stop limiting people to only use some platforms. So ArcGIS, they should not limit people to using ArcGIS so that they can open their data set. I think that for me would be it.

Maybe another thing, the intermediaries are increasing and they are duplicating data which sometimes confuse us, because subnational data is a big thing, boundaries data. I know more than 10 providers, intermediary providers of datasets. Sometimes somebody has come up with a with a small package that overlays boundaries from like 5 or more sources of the same country. So they are saying these are provinces, but five sources, they will not overlay. So then I wonder, why did these other intermediary come up to provide boundaries? Why wouldn't even they work together with other party to improve? They're starting to become many and they are starting to be a bit more confusing than helping. They are confusing people and adding layers of work because people will wonder, should I use HDX? Should I use Geoport or should I go to that country data? Then you end up with five types of boundaries that don't make sense at all. So those are the kind of things that I experience from time to time and I end up just using from one source and I don't know if that is the correct boundary. I know it reflects that country, but few boundaries will not match with another source's boundary. You can't tell and if you ask people around they will say they're not sure or something like that.

So I think the intermediaries should more work together to improve the data sets as opposed to getting more and more providers or more and more intermediaries in these game of data provision.

Interviewer
What about if you see a few sources that have like a different data, can you then see the metadata? Is it useful then to understand whether this is indeed from authoritative original source or even they say so, they are still different?

Interviewee
Yeah, even if you find the metadata sets that say this data was collected with the National Statistical Office of this country, together with the World Bank, for example, and you think, OK, this seems to be authoritative. Then other source will say something similar, not World Bank this time, maybe USAid. But still those datasets have the issues that they will not overlap, but they are representing the same thing. They will not overlap at all.

I think this data if they represent the same thing, they should be the same data set, but in most of the cases you'll find not, and that is especially for boundaries. Even if you go to population datasets, for example, World Bank produces datasets, you'll go to sites then from Columbia University, they also produce their own datasets. You go to another body, there are like 5 other actors in the population datasets. Of course we wrote a paper recently trying to compare all those datasets, they give you very different results when everybody's saying “our data is good, our data is good”. So you get more and more providers or intermediaries, but they tend to walk alone or even ignore the fact that there is an another provider or the same people that they can work with. Because they want money and they want to be in business, you will decide to go solo as opposed to improve whatever the other one is doing or even collaborate.

Interviewer
And from user perspective it's very confusing.

Interviewee
User perspective, especially for a researcher or even policymaker. A researcher can make logical decisions, but what about a policymaker? Because some guys from this country will come to Tanzania. Another guys from another country will come to Tanzania. They are presenting the same thing but very different outputs. I've seen where policymakers are confused on what we have here.

Interviewer
Yeah, well, that's a very good point about data confusion because of many intermediaries. So, not

necessarily many intermediaries would result in a good outcome for ecosystem, open data ecosystem. Very good point.

So that's the end of my questions. I would end my recording now.