COMPETITION COMMISSION OF SOUTH AFRICA

In the matter of

MEDIA AND DIGITAL PLATFORMS MARKET INQUIRY ("MDPMI")

held in hybrid format at

Dtic Campus, Sunnyside Pretoria and virtually via MS TEAMS

on 06 March 2024

Chairperson: Chief Economist and Acting Deputy Commissioner: Competition Commission: Mr. James Hodge

---

PANEL MEMBER: Ms. Paula Fray

Day 3: Med8 Media

Data Science for Social Impact Research Group: University of Pretoria

Mail & Guardian

START OF AFTERNOON PROCEEDINGS ON 6 MARCH 2024

CHAIRPERSON: Good morning and welcome to day three of the Media and Digital Platforms Market Inquiry. We have an interesting day planned. So our first stakeholder is from Med8 Media. We are then followed by our data scientist to assist us in understanding AI. And then in the afternoon slot, we've got the Mail & Guardian. So welcome [indistinct 00:10:38]. Thank you for making the time to come and speak to us. I think if we could get your presentation, but maybe begin with also a little bit about Medi8 and what it does and then your inputs to the inquiry.

MR. CHRISTOPHER MCINGA – MED8 MEDIA: Fantastic. Thank you so much, James and Paula. I think today I got it right, I didn't call you chairman. But thank you so much for this opportunity and also to the panel as well, thank you for affording us this opportunity to come through. Ja, we are going to be presenting today, but I think best is that I outline exactly who Med8 Media is and also just to give an introduction to myself, who I am and my history. I think that will best kind of frame the reason why I'm standing here or sitting here today. So let me start off by saying my name is Christopher Mcinga. I started my career in a community newspaper, Caxton Local Newspapers. And I worked myself up to a level of being digital right hand man to the then CEO, Bruce Sturgeon. I had to report to Bruce Sturgeon and also report through him to Piet Greyling, the late Piet Greyling as well. During my tenure at Caxton, I had the great opportunity to visit the UK during the time where they started realising that there was a decline in advertising revenue on the print, ROP and inserts for the newspapers of the Caxton Group. So they sent us to the UK to find some kind of solutions and to see other media organisations in London and other parts of the UK. So I visited publications or rather institutions like Johnston Press, which is almost like

a carbon copy of what Caxton is here in South Africa. And also Local World, which is also a community newspaper based organisation with a lot of small newspapers for the different suburbia in those areas. I came back more depressed than going there all excited because one of the executives at Johnston Press said, you see Chris, you see that area there, there used to be a printing press, but we had to dismantle it and sell it off for scrap because the audience has migrated to online. But we learned a lot of things when we were there at Johnston Press and we actually came back into the Caxton Group and initiated the actual digitisation drive of their print products. At the time, I'm not quite sure of the figure, I think it was more or less about 120 print editions or print publications across the country, of which we then managed through a team. It's interesting that Moneyweb was actually here yesterday because they also played a vital part in that digitisation drive, because we used to sometimes park at the Moneyweb offices and come up with strategies of how we're going to be able to deploy the digital transformation of print media for Caxton. So through a team we managed to, out of those 120, I think we managed to get about 70 plus of those print editions to have digital presence, to have news websites. At the time we launched something called Look Local and then eventually through a consultant called Gregor Waller, we managed to kind of move that strategy from having Look Local news sites to having title sites with the title of the newspaper. So ja, I then subsequently was redeployed, if I can call it that, to the capital city here in Pretoria to look after a company called Capital Media. Now the composition of that company, I'm not sure if I can actually share, but ja, so that was another beast that I had to kind of handle, having over 379,000 newspapers being produced by the Rekord Newspapers, which falls under Capital Media, delivered door-to-door, knock-and-drop operation on a weekly basis. So there was about nine titles that I had to look after. I then realised that my passion really lies

within township economy in rural areas and language, because I saw in the Caxton group there was not that much of a variety of publications that had diversity in languages because I'm really interested more in the vernacular, mother tongue, the township economy and what's going on there. And even during my time at Capital Media and Caxton, I started realising that a lot of advertisers at the time as well were not supporting the township-based newspapers. So that for me was a conundrum, saying, well, there's buying power in those communities. And at the time, I think there was this concept coined during that time when I was in that space called Black Diamonds and so on. So I decided, look, let me rather branch out and start my own thing, and let's see what I can do. Because my passion really lies in township, in rural areas, and development as well, entrepreneurs. So that's when we established Med8 Media. We registered Med8 Media, I think, in 2018. So it's a fairly new business, but I think, I don't want to call ourselves a small business, I will call myself a growing business, because we are in a sector or we're in an industry that I still feel has got fertile ground, not only just from digital opportunities, but also from the print perspective as well, despite it being so expensive. So that was my kind of thinking when it came to, we established Med8 Media. Now, Med8 Media is a media company. One of the divisions within our company is the publishing division. So we service independent community newspapers from all corners of the country. We are also publishers of our own product as well. Now, for the sake of this presentation, I just want to share a couple of disclaimers before I go into Med8 Media. I won't be mentioning any of our clients' names nor publications, because we're still scared of big tech. We don't want to be victimised. Just an antidote here as well. Having submitted on Monday through the entourage of [Sanev? 00:16:41], yesterday Facebook went down so I had a lot of people calling me and saying, Chris, what have

you done? So for a moment there, I really felt powerful. I really did. Honestly, for literally 10 seconds, I felt like Mark and Elon. And it got me thinking as well, to say, I'm just getting these kind of inboxes. People are DMing me. Imagine how much power they actually have on a day-to-day basis, waking up knowing that they actually have the communities at the palm of their hands. And for me, just that little moment, I felt like David having slain Goliath. So let me go back to Med8 Media. Med8 Media has a publishing division which looks after community independent publications. We render services such as digitisation of those products. Understanding that those products or those publishers come from areas where there's a skill shortage to do what we can do as Med8 Media. So we get a lot of inquiries or rather requests for quotes to help them with digital strategies. Not just building a website, but also having a social media platform, having a marketing strategy, etc. We also do the general services of those publications, like print layout and design of adverts, etc. So that's also part of the publishing side. As Med8 Media, we also have our own 100% owned publications within our own stable. Now that is 100% owned by Med8 Media. And then recently, last year, we embarked on a strategy of trying to help the smaller publications that are kind of closing shop because they would say, Chris, we're not printing, but we want a website. We'd quote them, then they'd say, we can't afford it. Then we'd say, well, if you can't afford it, it's a bit difficult. So what we decided to do then, we decided to just leverage our services for equity share of their businesses. So just a little 10% here, so that they can actually have access to our expertise, and we can build them websites. We can give them hosting. We can give them the domains as well for a little bit of an equity share. So I know it sounds like an ambitious approach, but I do strongly feel that we look up to people like Murdoch, we look up to people like Terry Moolman, etc. At the end of the day, we also want to become media owners in South Africa and play

a big role. And also, the reason why we embarked on that kind of strategy is that we also realise there's bigger opportunities for those smaller community newspapers, if they are represented by a central media company. From a fundraising perspective, going after philanthropic kind of funds, it would be easier to actually get them if we say, we have 30 newspapers under our stable, as opposed to each one of them going and looking for individual money. And also, from whatever the outcomes of these commissions are, whatever other opportunities that might come through, it's best to be actually, to use a company like Medi8 Media to be able to channel those opportunities to those publications because we have a vested interest in their operation. Now, the second set of clients that we have is bigger organisations. South African National Editors Forum is one of our clients. We are currently driving their community media digitisation drive, which is pretty much what we do as Medi8 Media as well, but with SANEF, they have also introduced community radio and community television. So we don't just build websites for these publishers, we also kind of deep dive into their business and look at kind of barriers that they are facing and try to find solutions for those. Some might, for example, we've got 15 partners right now which are in the actual project for SANEF, and it's going fairly well. And all of them have got different challenges. Some want to establish TikTok, not TikTok, podcasting from a newspaper, print newspaper, wanting to establish a podcasting station, and we're helping them with those kind of things. And other clients that we have is the Association of Independent Publishers, also one of our clients, and also the Foundation for Human Rights is also one of our clients, especially a project by the Foundation for Human Rights is called Masibambisane, which deals with GBV and deals with kind of area-specific initiatives around gender based violence, etc. So that's one of our clients as well as Medi8 Media. And I think I've mentioned pretty much quite a lot, and I have not

even moved on the slides. So I'm excited. I'm going to bring a different energy to this inquiry. And also, I just want to state as well that I'll be speaking the most simplest of language because in the past two days, I've heard a lot of jargon. I've heard a lot of kind of industry specific kind of concepts that can be complex for people that are actually affected by what we are here for. So by all means, I'll be speaking the layman's term. I'll be shooting from the heart. We do not have any legal representation when it comes to our presentation. This is really about reflecting on our clients' needs, and I think that's also very important to note. So this slide, I've kind of unpacked it thoroughly, and especially that italics part at the bottom there about a percentage ownership of smaller community newspapers. We don't want to see news deserts in South Africa, and we also believe that especially in the rural areas and the townships, there's still opportunities to establish newspapers, to establish credible newspapers in those communities. If suburbia is able to do it, why can't a small division of Diepkloof have its own newspaper as an example? So that's where the passion comes in. And also, I'd like to state that one of the 100% owned Med8 Media publications focuses on developing journalism as well. So we have MOUs with the university with regards to allowing their final year students that come out of university to do their work integrated learning with us. We do not get any kind of funding from SETAs, because it's very difficult to get that kind of funding, but we are actually leveraging as a company to say, look, we're giving back, and we want to make sure that the people coming out of the journalism faculty have got opportunities, because obviously, as we all know, there's a massive decline in number of editions that are available out there, for obvious reasons, which is advertising and also just the migration of audiences. But we do believe that as well, that there's a need to have these students practice journalism, and that if they are entrepreneurs, to aspire to having their own platforms, whether it's

online or whether it's print. So one of our publications fosters that. We mentor the guys. We give them the guidance. And also, recently, yay, we were able to actually absorb one of them into our team as well as an editorial administrator. So there we go. I think let me move on to the side. So who do we represent? We represent us as Med8 Media, as we are publishers as well, and also our clients. And I've unpacked our clients, small community media, independently owned companies and newspapers. And some of them have got a, I wanted to paint a picture of who they look like and what's the composition of their business so that it's clear. So we've noticed that some of them used to work for a Caxton or a Media24, and they branch out and they start their own thing. Some of them come from a sales background and establish a newspaper and get proper reporters to produce the publication. Some of them are one-man shows. They're a writer, reporter, distributor, layout artist, salesperson, one-man show. And some of them, yes, are hand-to-mouth operations, meaning that when they see these or when they receive a kind of a vacancy ad from a municipality, then there's a rush. We're going to print 5,000 copies. Chris, let's get it done. Let's put it together. We've got the content, we've got the articles. So they will put it together just to service that kind of need. And then some of them are really sustainable in a sense of printing monthly consistently. And also another thing that's very unique about Med8 Media is that we can see the wins from various smaller publications. So whatever win from one publication, we can then also kind of share with the other one to say, look, they've established a good relationship with the municipality and this is the angle that they use therefore, you should try and do the same. And also show them your publication to see this is the kind of advertising that they're placing in the printed newspaper. But had Med8 Media not existed, they would be still working in silos and only believing within what is in the environment, and probably news deserts being established there. And

then also we partially own some of the publications, and also our clients, like South African National Editors Forum, we are driving their digitisation drive. So that's pretty much what this slide is talking about. So we are submitting three items, and very basic three items. And we decided as a team that we don't want to go and touch on too many things, like Facebook and TikTok, let's just focus on Google. Let's focus on something, because we know our colleagues, our big brothers and cousins, other media companies, will probably do a more comprehensive job in sharing their experiences. And also before I start talking about what we are submitting here, is that I strongly believe that we, as the community media space in South Africa, and I'm speaking on behalf of our clients, and by no means am I saying all the other people are the same, I believe that we are at the right place at the right time, that we're having these conversations now rather than later. Rather than not have these conversations and looking back and saying, we wish we could have had these conversations. So we come in good faith, and we want to welcome the opportunity of just being in the room with big tech. Not even just having a seat at the table, just being in the room and having a comment, a cough here and there, to say, oh, hold on, don't forget we are here. So having this opportunity pretty much is massive for us. It's a massive marketing opportunity as well for Med8 Media and all the other media companies out there that want our services, please feel free to contact us. Alright, so the first one is really the discrimination against indigenous language news sites. Now, what we mean about this is that, you know, aside from Google not prioritising community news, we also feel that there's a discrimination when it comes to the indigenous language websites. I understand that they've got a Google news policy in place, and they allow the crawlers to pretty much crawl and feature news. However, there's a lot of kind of work that needs to be put in, especially that of having constant publishing online. So if you're a

new guy, if you're a new publisher, and you are publishing indigenous language or just perhaps English, it's going to take you a long time to actually start featuring Google news because I do understand they're talking about things like authority, things like, you know, the expertise and the trustworthiness, but that can only be done when you have got a lot of consistency. Now, remember, I painted a picture that, you know, these publishers, some of them are one-man shows. So I recall having to apply some kind of change management strategies in the newsrooms for the Caxton Group to say, look guys, you know, you used to be just writing things down. When you're out on a scene, in a story, try and tweet as well, try and Facebook. And try and ask for people, what they're thinking of the strike, whatever the case is and try and develop that story. And yes, you can go back to the office and write a comprehensive article about it. But know that it's a living thing. And don't forget to take a photo and maybe a video here and there. So now you're kind of having multi little, you know, but now imagine this for an independent newspaper owner who has to do all of those things and also upload his own articles online and tweet and Facebook, etc. So it becomes really tricky from a workflow perspective to kind of get them to publish all the time. So what we are doing with the digitisation drive for SANEF is that we're saying, guys, set yourself targets, know that you need to create an expectation from your clients that if they go to your website, at least there needs to be something fresh every single day. And news, if it's not happening, news, question Paul asked, what is news? You know, make sure there's content, let me put it that way. Whether it is soft fluff that is on the website, but also news. Just make sure that there's something new on the website all the time. So going back onto this discrimination, and I'm using a very strong word there, discrimination, I do feel that it feels like we are being discriminated against, especially on the basis of indigenous languages. And those sites as well, you know, they serve

as archives for language. They serve as archives and access to information for the children, the future. So if we are not even prioritised or discoverable online, it means that now that we can't print, our kids are not going to have any kind of properly written in indigenous language material. It means that we are going to be losing languages. It means, I recall growing up, I used to read the Sunday Times with my dad. I used to read that little magazine with Arthur, you know, that cartoon thing in the middle, I don't know if they still have those. But I used to read, and I used to go to school, and they used to give us an assignment to go and read about a certain topic at the time. I remember one very well of Bob, our neighbour, late Bob, in Zimbabwe, during the time of the farms and etc. So that was, for me, very interesting because I was into politics, and I'm not there now because I'm too soft for that. So I used to have access to information through the Sunday Times, and it was written in English because I went to an English school. Now I'm looking at it and saying, what about a guy sitting in Giyani who is going to a school that is taught in a vernacular language? Are they given assignments to go and read literature or read content? And if they do, where do they find that information? Do they go to the library? Do they go to the school library? And so what I'm saying is that the community newspaper plays a vital role as well as reading material. So if it's not discoverable on Google, it means that we are actually going to be killing the language, which is coupled by culture as well, closely. And in essence, we are saying our languages, our indigenous languages, are not important, not really prioritised by Google. I don't want to go too much into this because this is really emotional for me because I strongly feel that we are going to be losing our flavour as South African, our authenticity, and our culture in communities. So that first submission, I will kind of talk in general about it in other slides, and the second one is that of AdSense. Now, did you know that Google will not allow AdSense or display ads

to be visible or to be shown on indigenous language news websites? So it means that we've got no opportunity whatsoever to ever get a little taste of digital ad spend. So we are excluded from that conversation. Yes, I do understand that they have about 49 official supported languages that they allow AdSense on. And for South Africa, it's only English that I'm aware of. And unfortunately, that kind of leaves a bad taste because are we saying that credible news sites that are of authority, that are trustworthy, that are prioritised need to be in English in order for it to kind of attract buying, buyers from the advertising? I don't know. I really do strongly feel, and like I said, I'm shooting from the heart, I'm talking layman's term here. What does that look like for an aspiring publisher coming out of university or having served for one of, or having retrenched, let me go that direction, having been retrenched from a big media organisation, and they want to start their own newspaper, and they want to start it online because that's where their traffic is, and that's where their audience is. And are you saying if they write in indigenous language, they're not going to be able to monetise? Look, I come from the school of hard knocks, pounding the pavement, selling advertising door to door, entering a panel beating shop and waiting for that person to finish doing a quote before they come talk to you. So we know how to sell direct advertising from clients. That cheese has moved. It's moved now to digital agencies within communities. We used to be able to get that money for the printed newspaper because we had something tangible to show, and they also were reading it. Now let's paint this picture as well in terms of just the advertising space. If you look at some of the businesses, they're established by someone or started by someone, and that person becomes old, but that person used to grow up with newspapers. Now they hand it over to their children to handle the business. The children now are more inclined to be on cell phones. So their advertising spend is going to shift towards where they are. So that's

just the reality. That's just what it is. There already, the cheese has moved, and we need to find new cheese. And to be excluded like this, just because you are a publication that is written in an indigenous language, it means that you're not going to feature at all. Unless, obviously, you go through a company that aspires to also have ad service like Med8 Media that's going to service this kind of market. That's what we're aspiring to do as a media company. So AdSense not being featured by Google, for these languages, I feel it's discriminating. I understand maybe from a business perspective it's impossible to hire more people, to vet all the content, etc. Maybe it's going to be cost on their side. But surely there should be kind of leniency towards countries like South Africa when it comes to just the ad spend. Now on ad spend as well, look, at the end of the day I know that we're not going to be making thousands and thousands of rands because AdSense comes with the traffic. That goes to the next submission that I'll be making now. But with AdSense, we just want a taste of digital revenue for publications of indigenous languages. Because we know the publishers that we service that are English, they do have AdSense. However, they're not generating that much money. So that goes to my third submission here, is that those smaller English newspapers that are running AdSense, there's a high threshold. Before you can cash out money, you need to reach at least R1,000 and then obviously it's all about the numbers game. The more traffic that you have, the more likelihood you're going to be generating money and it's not going to be a lot of money. Now I'm not saying that Google need to kind of put the bar down, I'm saying perhaps the CPM model needs to be looked at and maybe prioritise a higher CPM that will go to publishers as opposed to such a small amount because we're not playing in a numbers game in a community space. Let's look at it in the context of a small community, let's say 20,000, 25,000 population. Let's say half of them, from a literacy perspective, let's

halve it like this. Let's say 25,000 and then we say half of them are most likely not going to be online at all because of the cost of internet connectivity, etc. Now you're half of that 25. Then you halve it again based on literacy and likelihood to want to read news. Then you halve it again because of news avoidance. Then you halve it again because of age group. So now you're looking at maybe just generating 2,000 page views or unique visitors or page views per site. So you're generating a lot of money from that is highly unlikely going to happen. So I don't know what the answer would be, but I would really like to say that we have to look at it from a different perspective. Let the content that we are producing, especially that in indigenous languages, become more expensive because the cost is high of producing quality journalism. They're real hard costs. Now I'm speaking from a layman's perspective. The first one being fuel price is high. Just living is high. The cost of connectivity to do your research is high. Then there's also you need to eat at the end of the day. That's high as well. You need to live. So in producing content, quality journalism, and it not being that much appreciated from a monetising perspective, I'd say that it's a bit disjointed. We need to value our content higher than just generative stuff that's being produced now. It seems as if it's the one that's been prioritised more by Google. They're pretty much double dipping in a sense, prioritising other stuff online and deprioritising the real content. So I've spoken about these things in depth and I've actually got presentations on each one of them, but just for the purposes of rehatching it, I'd like to say, look, Google does not appreciate indigenous language websites. It does not make it discoverable. They've got all the tech with devices that come automatically loaded with Google Chrome and Google Discover. Surely, geographically, with the pin, they could sense that that town should have a community newspaper and that news should be prioritised. I'm not a tech genius here, but I'm just saying from a simplistic point of view,

if you are able to target me, surely you'll be able to target me with local news as well. So that's number one. And then the second one is AdSense, just to reiterate, 49 supported sites. And also, this was a policy just recently published, I think in 2020. And obviously, on that list, English is the only one. And I repeat again, it feels as though in order for you to be a credible news source, you need to be writing in English. For me, it's a swear word, because the thing is you're saying that a journalist that's writing in an indigenous language is not credible. You can look at it both ways. So that's what I'm submitting there. And then also on the AdSense side. So AdSense, just to break it down, look, this shifts regularly and differs from country to country. Just to give our publishers and also our clients an opportunity to understand this, you'd need to generate pretty much 10,000 impressions just to get that R1,000. Now, 10,000 impressions for a community, like I broke it down just now, is most likely going to be generating just 2,000 page views. And obviously, you can put more ads on that page to get those impressions, but it'll make your site look like a gambling site. So ja, just to give you a better understanding of how we're thinking about this. And also, just on this slide, I failed to mention this, let's just value, especially indigenous language sites and publishers, let's just value their content more because they're preserving a language, they're preserving a culture, and they're also improving on the literacy of that particular community. And we should not be forced to eat English as being the only credible thing. So ja, in conclusion, Google pretty much is saying to us, look, you're not important, only English websites are important. We'll prioritise them. Secondly, Google is just saying, look guys, you're too small, indigenous languages. We don't even want to give you any AdSense money. You don't deserve it. That's how we feel. And thirdly, again, the bar's set high and the likelihood of us actually generating any revenue online is very bleak. And I stand here also representing my clients through the SANEF

digitisation drive. It would be irresponsible of me not to actually have had an input into this because we are trying to make sure that they are sustainable. We are trying to make sure that they do not die down. That is why we help them with the digital transformation of their traditional media. So that's pretty much me at the end of the day. I'm more than welcome to take any questions and hope as well if there are any other questions that you asked in previous days that I can actually maybe give more simplistic answers to for the sake of the people that we represent. Thank you.

MS. PAULA FRAY: Absolutely, thank you very much, Chris. That was really a good start to the conversation. We've got a couple of questions. I mean, just to get a clearer understanding of the unique challenges that community media are facing in the digital world. Won't you tell us a little bit more about the SANEF digitalisation project, what it is and why the project was started?

MR. CHRISTOPHER MCINGA – MED8 MEDIA: I thought you were going to shoot a couple of questions. I'll address all of them. Are there any other questions? I sound like a politician now. Alright, thank you so much, Paula, for that question. It's a very important question and it's also a question that will help me market this initiative. It's not a project that has got a bottomless funding. It's a project that's specific and that is why we could only take 15 publishers onto this drive. We had over 68 applicants for this and we could only choose the publications and also radio stations that we've chosen. So the composition of that 15, call it a cohort or call it beneficiaries of this project, is that they comprise of a women based magazine. They also comprise of community radio stations which also, through ICASA and everything, they are live on air, but then we have to kind of apply other digitisation mechanisms like having them stream online to Iono FM, which was mentioned the other day as well. So those are the kind of things that we are implementing with them. And then we tried to get some

community television in, but the applicants that we got were not from community television, they were actually start-ups with YouTube channels, but they decided to apply. So we had a couple of those, but obviously for the benefit of the project we thought let's rather focus on the tangible kind of businesses like your newspaper, your magazine, your radio station, and if you had a community television like Soweto TV as an example, if they had applied, we probably would have selected them because we wanted to see how they can actually also participate in the digital space. So the drive itself, I cannot speak on behalf of SANEF as to why it was started, but I can assume that it is to counter the decline of credible news platforms from a community newspaper perspective. They are shutting down, so we need to kind of move where the cheese is per se in terms of audience migration. So I think SANEF's thinking here was also to say, look, how can we save these publications, but also make them sustainable, self-sufficient? They can supply for themselves, and that's why I'm here today, to say, look, hey, it's a difficult question to ask because I drive a hard sell when it comes to our partners. I tell them straight forward, guys, after this 12 month period, you will pay for your domain by yourself. You will pay your hosting for yourself. You will still continue to produce news for yourself. So from day one, start making sure you start monetising and trying to find money, and then we help them as well, not just to build websites. We build the websites for them, we teach them how to run their website and how to market their website. Some have got aspirations to setting up TikTok accounts. Some have aspirations to set up podcasts that we best advise them of the best practice and how to do it. Obviously, from our side as well, we are learning as Med8 Media. From this project, we will be able to take on any other project as well and replicate the wins of this particular project. So I really commend SANEF for actually initiating this. And then the other part of this drive, Paula, is that we also established a community media

support services brand and website, community media support services. Now, community media support services is a website whereby all the other participants, or all the other applicants, rather, that were not successful for this drive, they can go onto that website and have access to tutorials on how to publish online, how to market your website, have resources, presentations that we give the guys that are in the project, but they also have an opportunity to participate in there as well. And we also have a paid version of the community media support services, which is available for anyone who wants to participate in the digital media landscape. We also have a paid version of that support services and for a fixed fee, they will get reduced costs for things like hosting, things like domain registration. They will have access to our developer at a reduced cost because if you had to sit down with me and you say, Chris, give me an hour, that's a consulting fee. It's your business, it's not mine, and you want my help. So we give them access to those expertise as well. And there's other things like being invited to training. So we're having a bi-weekly Monday at 10am training where we invite via Zoom an expert in a specific topic and then they'll have access to pop-in training to gain that insight. So ja, so that is the SANEF digitisation drive. And again, I want to emphasise, it's not just about giving them a website. It's really about deep diving into their businesses and unlocking, and unlocking things like, is your rate card sellable? Where did you derive your rate card from? Some publishers base their rate cards on the competitor or they base their rate card on someone else's rate card, but they do not know the science behind the per column centimetre or the science behind yields in a publication, the science behind advertising versus editorial space and also keeping your publication, and I'm talking now in particular the print publication, keeping it away from being a catalogue and rather being something that has got value. So we'd go into things like that and also did the development of digital rate cards as well. And

now I'm talking about directly sold banner advertising. So that's pretty much what we also cover in this digitisation drive.

MS. PAULA FRAY: I can understand the shift away from legacy media because of where audiences are shifting, but given the size of community media, Chris, I'm wondering whether you have determined whether there is a sustainable path for small media online given the current digital market and also the platform rules?

MR. CHRISTOPHER MCINGA – MED8 MEDIA: I'll attempt to answer that question, Paula, because I think you're bringing the concept of size of the communities and also the sustainability trajectory within that space. You know, again, I'd like to emphasize the community media space is still fertile. There's still some space for players. Whether we are small parts of a bigger picture, let's rather have those small parts of this bigger picture. The more we can have new platforms also participating, maybe then there would be a need from, let's say, Med8 Media to establish an aggregator that's going to prioritise those smaller little parts. So for us, we're looking at it as Med8 Media from a long term perspective saying, look, if we support these guys and we maintain those smaller ones that are existing and maybe we get more established, we then can kind of create our own ecosystem and prioritise our own content through an aggregator of some sort. So yes, we are trying to play within that space. And also we see commercial opportunities in that space as well. For example, by June this year, we should be establishing our own ad server, well, white labelling ad server that exists, and servicing ads through these community media websites that we are building. And we see that as valuable because the likes of the GCIS can use us as a distributor of government communication through our ad server. So that kind of outlook is what we are looking at. Yes, we know that we stand no chance from sustainability point of view with them operating in little silos. But as a collective and being smaller parts of a bigger picture,

I think we still can keep, especially indigenous languages out there, credible news out there, community-based, targeted, hyper-local content. I hope that answers your question, Paula.

MS. PAULA FRAY: I mean, to talk a little bit more about your own personal experience moving from print to digital. I mean, I understand that when you shifted, that a lot of your print advertiser didn't necessarily shift with you but went directly to Facebook and Google with the advertising. Could you speak a bit about that?

MR. CHRISTOPHER MCINGA – MED8 MEDIA: Sure. So I'll draw you more experience from my days at the community newspapers, Caxton. You know, we saw that, well, I saw in particular, because I had to establish a digital department, one of the first digital departments in the Caxton group, which then was replicated across the provinces. Yes, there was a direct migration of traditional ad spend to going online. And it was not necessarily that the business owner would just whip up their credit card and do an ad campaign. But it was more so that they would use a local digital company, digital marketing company, to do their advertising. So, yes, we would have to, many times, come across that kind of objection, saying, look, guys, I know how many people I'm getting if I put X amount of money. And then, Paula, I'll just segue onto that as well. Because, you know, as part of us sharing our pain points, I also want to share some thinking around that as well. So if the likes of Facebook, and I thought I would touch on Facebook, but if they are able to project how many people this boost is going to reach, surely they can put value to that. You know, so the maths is there. It's just a matter of how do we then compensate the originators of that article, that content, and channel that revenue to them, and them then taking the agency commission. So I'm repeating agency commission, because I just think this is so simple in my head, that they're just an agency. The originators of the content need to get the bulk of the money,

and then the agency, which is the distribution channel, needs to get the agency comm. If they work hard, everyone gets paid, and they get more money, and they can do it. So anyway, so let me go back to your question. The drop off was mainly from smaller businesses, and not so for the bigger guys. So if you look at the community newspaper print business, you would have seen as well, within your communities that you live in, that there's been a kind of a pinch factor when it comes to the pagination of each publication, and more so that there's now more inserts than there are ROP newspapers. So I don't want to use the swear word of community newspapers becoming now carriers of catalogues, but it's evident, it's there. So the smaller guys who used to advertise and play a big role in bumping up the publication through features, specifically around Valentine's Day or Black Friday, they would be on ROP. Now that has migrated, for the big guys, to inserts, but for the small guys, it's migrated now to an ad agency, buying on their behalf and getting placed somewhere else. So the big guys still see the importance of keeping their so list or their earpiece or keeping their front page ad. They know because they see the results. So even though the carrier itself is having more inserts, they know the people are going for inserts, but you will still have that newspaper in your house.

MS. PAULA FRAY: Chris, I know that you wanted to focus on Google, but could you speak to how important the social media platforms are to small community and vernacular media, please? Particularly Facebook and WhatsApp?

MR. CHRISTOPHER MCINGA – MED8 MEDIA: Sure, Paula. I was steering away from Facebook because it's a love-hate relationship. If I was a Facebook profiler, I'd say it's complicated. Our relationship is really complicated. I feel entangled about this because we help the newspapers set up Facebook pages so that they can drive traffic to their website. We help them not only to set up their pages to drive traffic, but also to

encourage engagement on Facebook because that's where the people are. At the same time, we also encourage them, yes Facebook, I will say this, we encourage them to also sell directly to your local car wash that, you know what, that poster, we'll place it on the website and we'll also place it on Facebook as well and we'll pin it at X amount of money. Yes, we are deriving as well some small bits of revenue from placing those pamphlets on Facebook even though it's against their policies somehow. So those are the kind of things that we do for the smaller community newspapers. The relationship is a love-hate relationship. They derive majority of their traffic through Facebook and then we see also some of the publishers are starting, which is against our advice, they're starting to publish the whole article on Facebook and keeping the people there, saying, Chris, we're getting more likes, we can see directly that it's happening. So there's a little bit of education that needs to be had there as well and also we just recently started encouraging them to put their links in the comments below.

MS. PAULA FRAY: So the audience is on Facebook more than on the website?

MR. CHRISTOPHER MCINGA – MED8 MEDIA: The audience is more on Facebook than the website and also touching on WhatsApp, we've seen quite a bit of, you know, I said I won't mention a particular publisher, but there's one particular publisher that's doing it very well. So they're utilising statuses as part of their marketing of the news. So yes, WhatsApp statuses also yield some traffic to go to the website because once the people save your newspaper as a contact on your phone, you are going to see their statuses. So they are playing within that kind of attention economy as well on WhatsApp. And then also on WhatsApp, they distribute e-editions. So there's a lot of publications that send through the print ready copy to your printers to get printed and then they create a more compressed version so that it can be shared amongst people via WhatsApp. So that's also something that we've seen work. But whether the people

are actually, because I, myself, I just look at it because it's art coming from, I've got ink in my veins if I can call it that, but it's art. So on the mobile device, it's not really that reader friendly but at least if the layout looks amazing, then it will be featured in the newspaper whereas if you feature it on a Facebook post or on a website, it's kind of it's there, it's not really seen.

CHAIRPERSON: Maybe I can just pick up on a couple of those themes. So we heard yesterday about the whole pay-for-news type arrangement, but would your relationship be less complicated with Facebook if you had more monetisation opportunities on Facebook?

MR. CHRISTOPHER MCINGA – MED8 MEDIA: Yes. I think in a relationship, James, you can't always be receiving, you also have to kind of give. So in this particular case, Facebook is receiving. It's not giving. And it's an entanglement because it's an abusive relationship, for lack of better words. So yes, if there were opportunities to monetise or have a revenue share of some sort, then I think we don't need counselling, marital counselling. But ja, I think in particular, if they were forthcoming to sharing a little bit more and having these kind of discussions, we might come up with a makeable solution.

CHAIRPERSON: And given your experience, I mean, they obviously have their business model, but coming from your side, now you're in the room at least, where do you think you could monetise more on Facebook that you're being stopped at the moment?

MR. CHRISTOPHER MCINGA – MED8 MEDIA: As a publisher. As a publisher. Look, I'm wearing different hats, right, so I'm going to talk as a publisher, not as a media agency and a media publisher.

CHAIRPERSON: Absolutely.

MR. CHRISTOPHER MCINGA – MED8 MEDIA: So look, monetising opportunities would mean that, especially for Facebook, would mean that if they could make it available, that they offer our clients as a publication an opportunity to work through us rather than to work through independent agencies. So for example, I just recently came back from, recently, a couple of months back, from a sustainability conference, a sustainability and journalism conference in Nairobi, Kenya at Aga Khan University. And the key takeouts out of that kind of discussion with all the other media in Africa was that we need to, as news organisations, as media organisations, we need to have an octopus approach when running our legacy business. Meaning that we need to have multiple revenue streams coming through and not just from advertising. So for example, if it's goods and services that you need to have as monetising opportunities, then why don't you sell cups and T-shirts, whatever the case is. If it were to be a newspaper but acting as an ad agency or marketing agency on behalf of your clients in order to retain them, why can't we just then help boost their commercial opportunities, their adverts, but as a newspaper? Yes, we can do that now, but if Facebook can, in a way, showcase that it's better to deal with your local newspaper or your advertising, as opposed to saying, give us a credit card number, we can reach so many people. So I believe that we can service our clients, our direct clients in the communities much better if Facebook can allow us to be the agencies and not just make it freely available for people to boost their own. I know it's very opportunistic of me, but I just want a little bit more hand-holding from Facebook in this relationship.
CHAIRPERSON: Well, look, I don't want you to have an affair, as we've got to go there, but it is interesting in looking at it, YouTube has a model where they have partners where if you're a partner, you can sell your advertising direct through them, which also means, presumably, you might be able to get more targeted and better cost per

thousand. So it may not be completely new. I mean, related then to Paula's question on where's the audience. So one can look at where the revenue is and say, oh well, the audience seems to be on Facebook or wherever, but you've raised the point, well, I'm battling to get visibility on Google. So you can read the statistic, you don't get much from Google because everyone's on Facebook or because you can't get much out of Google. So my question is, where is the actual audience? Are they on the web using Google? Are they on Facebook? So even if Google fixed this, I mean, is there a revenue stream there?

MR. CHRISTOPHER MCINGA – MED8 MEDIA: I'd say categorically that the audience on Facebook, they're on social media platforms, if I can just generalise in that way. Yes, Google is there for search. For some people, it's really to find how to do things or do homework and research. And pretty much we're always saying that Google is pushing news to you, I mean, through Discovery and Google News. We're not featuring at all. But the audience is engaging on Facebook, and some of them, you know, I'd like to say that there's a need as well within this space, and especially the community media space as well, for news literacy. I know I'm segueing into another topic now, but we need to help our readers as well to discern the difference between credible news and disinformation and misinformation. So that's lacking already there. So there's dangerous playing ground there on Facebook, and that's why I don't want to touch on this conversation, because it's so complex. It's very convoluted. So the audience is there. They need to be taught what is really credible news and how do you fact-check news. And some of them, I would like to say, and I've seen this a lot of times, where we publish something on a Facebook page, it gets shared, and people are just commenting based on the headline. They haven't read the article and we're like, wait a second, this comment seems like they just read the headline and made up

their own conclusion and not actually diving deep into reading. So I'd say the noise, the club, is on Facebook. That's where the club is. The restaurant is on Google. So there's more people on the dance floor here and very less people sitting down that are susceptible and willing to read news. They're sitting in a restaurant, a very quiet environment. So ja, Facebook has got all the audience. Google, maybe if they were prioritising our content even more, they would be there. Maybe if they start waking up, sorry, my readers, not meaning that you're sleeping at all, but as general as society, if we start waking up and saying, look guys, news is affecting your mental wellbeing and that is why there's news avoidance as well in the picture. And once you don't want to read and block all the news, you might want to get it somewhere else. And so if Google prioritises languages, prioritises geographic community based newspapers because of interest of hyper-local content that means something to those people, written by someone with a face, that they know that this is written by that journalist, or that my child's photo was taken at a graduation or a matric dance, and it's featured in the local newspaper, they can focus on that and put it up and show it up on Google. I think Google would then be a household name rather than a device. I think they're just a vehicle, not a household name. I know you'd say you Google things, but I just feel that they'll have a stronger bond with the people that are using it rather than it just being a tool.

CHAIRPERSON: And then I just wanted to pick on something you'd said. I mean, we started on day one and you were here with also questions around what is news. But your foundation for human rights is focused on gender-based violence and...

MR. CHRISTOPHER MCINGA – MED8 MEDIA: Especially the Masibambisane initiative.

CHAIRPERSON: And I mean, that is educational. So it might not fit into, everyone has

this view of news, or maybe not everyone, but news being the sort of political events, the big events, the national events. You've now mentioned having your daughter's picture from the matric dance in the local newspaper. So for your communities, what are the important things that you deliver that are not being delivered elsewhere?

MR. CHRISTOPHER MCINGA – MED8 MEDIA: Very good question. And I think that's what differentiates us from mainstream media. It is that local touch. It is you knowing that journalist, carrying that camera at the rugby match. You know that if your photo has been taken, it might end up in the newspaper. I've been in the Graaff Reinet Advertiser so many times for my rugby and my golf. So that to me is worth clipping out and putting into, I don't know if you still have albums, but clipping out and keeping in the albums, right? So in that sense, whatever I believe is worth keeping to that particular person is news. Or whatever I can talk about with my friend at the tea party that I saw in the newspaper, I believe that is news, community news. Because yes, it's not news, hot news covering for W's and H. Obviously most of them need to have that. That's the foundation, right? But I think if it's worth talking about, then it's news. You know, so yes. Let me answer your question straightforward. Anything that's relating to problems in the municipality or issues that are affecting us, that's news in the community space. Anything that is breaking, yes. Even though they can't break it immediately like a daily newspaper would. But if a house burns down or if a shop burns down, people saw the smoke, they need to know what happened and they can get that from the community newspaper. Thirdly, yes, it's school news. Meaning your little Johnny's achievement, spelling bee or whatever the case is, that is also news because I can talk to it and brag about it in the brag pages, right? And then also, something that is dying, if not dead, it's the classified section. You know, obituaries and also notices. And also the plumbers and the electricians that feature there. I know it's not news, but

it's important as well because you need to find it somewhere. When that pipe bursts, you need to know where to go. And the local newspaper is usually where you would go. So, yes, current events that happen and also entertainment as well. You know, what's happening, what's on, what's coming up. Shorts and shades, events happening. People want to know where it's happening and also once it has happened, they also want to see how I featured in that paper or what happened as the post-event kind of reporting. So those are the kind of fundamental pillars of the composition of a community newspaper.

CHAIRPERESON: Thank you so much, Chris, for coming. I think it has been very insightful. I wish we had more time, but we are following up with questions and getting insights from people. And if there's also anything you feel that we didn't cover that is important for us or want to flesh out on anything you've said already, then please also just make a submission. But I think it is an emotional issue, as you put it, and there's no shying away from that, I'm afraid. And that's fine.

MR. CHRISTOPHER MCINGA – MED8 MEDIA: I'd like to just conclude and say last year through the Association of Independent Publishers and the American Embassy, they afforded us, ten publishers, an opportunity to spend some time at Ohio University down in Athens, Ohio, in USA. And we went there very much expecting to see tech and to see things around digital and technology, etc. And you know what, we were faced with the realisation that newspapers are not dead. Newspapers are still thriving in these communities, especially in the Ohio State, because we were only exposed to the Ohio State. So there's still space for the printed newspaper. You've got publications like the Budget, which focuses mainly on the religious group within a specific area. That newspaper is thick. I couldn't even put it in my bag because I probably would have been overweight at the airport. But the newspaper space is still thriving in the

United States, where we were exposed to. However, they do not neglect digital. It goes hand in hand. The other thing I'd like to state as well is that the advertising support from those communities is still there because of the involvement of that particular newspaper in the community. So they are part and parcel of being involved in a community event that's coming up. They're media sponsors in a community event. They are also part and parcel of solutions journalism as well, things that are around not just reporting on the problem, but also finding evidence elsewhere and showcasing within that community to say, come up with your own solutions based on the models that we have actually sourced. Solutions journalism is something that we need to also focus on because it really is going to be, I believe, a saving grace to a lot of communities because it's evidence-based, and also it's inspiring to come up with those things. If people are not exposed to that, then they're never going to solve their own problems. And then another thing as well is the trip to Ohio University also helped us come back with more ideas on how we can transform our organisation, Med8 Media in itself. We came back and we said, look, we need to restructure. We need to apply this and apply this. So all newspapers that are listening in today, you need to be agile. You need to be receptive of multiple revenue streams. You're not just in the business of news. You're also in the business of sustaining your business. So whether you are wanting to go after donor funding, whether you want to ask your audience for donations, whether you want to start up newsletters for email, email newsletters, try things out, share the learnings. And there's no one-size-fits-all in this situation, there's no silver bullet whatsoever. But if we can just learn from each other, that will really help. And I think institutions like the Association of Independent Publishers and also South African National Editors Forum through digitisation drive is also aiming to allow people to see more information, see more opportunities so that they can sustain their

businesses. So thank you, James, and thank you, Paula, and thank you to the panel as well.

CHAIRPERSON: Thanks. And we're going to take a few minutes break while we set up for our next stakeholder. Alright, we are back for the morning session, and our second stakeholder is Vukosi Marivate, the ABSA Chair of Data Science and Associate Professor of Computer Science at the University of Pretoria. Welcome, Vukosi.

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: Thanks for having me.

CHAIRPERSON: And we had specifically invited Vukosi partly to educate, I think, ourselves, the technical team and the public, on AI, because it seems a bit impenetrable for most of us. But I'm hoping after this, we'll be a lot more enlightened. So thank you for coming.

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: Alright, thanks. So do I go on, or...?

CHAIRPERSON: Yeah, you can certainly start your presentation and talk us through. We've got two hours, so we have plenty of time.

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: Alright, thank you for this opportunity, and also for the introduction. Yes, I'm at the University of Pretoria, running the Data Science for Social Impact research group or lab. I'm also a co-founder of two grassroots AI organisations that work to really make sure that African AI is something that we champion and make sure that Africans shape this ongoing revolution, both in AI in general and then in language, more specifically when it comes to Masakhane. I also do spend some time at Lelapa AI, which is an AI start-up based in Johannesburg. And I sit on the board or steering committee of the Lacuna Fund that funds the creation of data sets for AI, especially across the global majority, or global south, for lack of a better word, inside there. That's our lab at the university, made up of many staff

members, senior members, students. And about 70% of the work that we do is in the intersection of natural language processing and African languages, and the others is data science and society. And I hope by being here, I am also fulfilling our part in data science and society and making sure that we can form as a voice that people can trust. Given that we are a university lab, we do have some support from different organisations, as was said at the beginning, one being from ABSA for my position at the university, and then a number of research grants and gifts that we get from a variety of partners that work with us, whether supporting students, supporting post docs, or supporting infrastructure that we need to go ahead and do our work. But yet, I think to maybe say, given this, we do value our academic freedom within this and being able to do what we do more for the benefit of society in doing our research as opposed to sponsors or people. What are we going to do in this talk? And as I said, thanks again for the commission for having me, the technical team, and then also to the public, is we want to build some understanding of fundamentals. I know for many, if we think back, it's the 30th of November, 2022, when a large language model became public and people started using it. It was a flashpoint, but there are things we can all get to understand jointly that get us to really know some of the things that happen in the back end. And as such, then we need to demystify and then share some food for thought in terms of thinking further than just this talk. So to kick us off, let's first start with actually what's AI and machine learning before we get to the fancy things. So we're going to go from zero all the way to actually some of the services that you might be using on your phone or laptop. So what's artificial intelligence? In a classical definition, you have an agent. It exists in some environment. It takes action to achieve some goal. The humble South African robot lives at a corner of a street. That's its environment. It can turn from green, orange, to red. Those are the actions that it takes, and the goal is to

control traffic. That's an artificial intelligence. That's just to show that we build on from there. You can have a robot vacuum cleaner there. One of my twin daughters is looking at in a house. The house is the environment. It can move around. It can vacuum. It can avoid obstacles. Those are the actions that it can take, and then the goals are to keep the house clean. That's another artificial intelligence, and you can now start kind of thinking about systems that are like this everywhere. What's machine learning? Machine learning is a subset of AI that deals with learning patterns from data. You will hear these normally interchangeably used, but in just setting up definitions, we learn patterns from data. Right now, we're going to talk about supervised learning, specifically classification or categorisation. So let's say I give you a surface. It's a flat surface here, and then I give you two data points. Their positioning is important, but then also is their colour. So I give you two data points, one being red, one being blue, and I ask you in the future, if I drop a data point somewhere on the surface, let's say here, what colour would you predict it to be given that I've given you those pieces of data before? What you need to do is you have to learn some kind of discriminating function, and the easiest one, given the two data points that we have here, is just a simple straight line that splits the surface into two. It says, anytime in the future, if something falls on the left of that line, I'll predict that it's red, even if I don't know what the colour is, and if it falls onto the right of that line, I'll predict that it's blue. Right, so that we've just built. That's our first kind of discrimination function or classification function. Yes, life is not that easy. Typically, you don't have things that are just as easy separable. It tends to be that your data is more complex. Because your data is more complex, the pattern you have to learn also becomes more complex. So you see now, you've got almost like a quadratic equation or line that is now squiggly, instead of just being straight. Instead of just being Y is equal to MX + C. In this new one, anything

that's above the line in the future, I'll predict that it's blue. Anything that's below the dotted line, I'll predict that it's red. So you can see that you can go from data, and then you can learn this discrimination function, and it's useful for doing predictions in the future. So this is classification. We can take a leap, a semester or two in AI, and say this is how we get to, you have systems where now, you can take an image that you're seeing there, and then push it through these, what you call deep neural networks, which is really a parametisation or that discrimination function. That's what that deep neural network has, is that function, and then it tells you at the end of the day that what you have in that image is a dog and not a cat. That's what's happening. We go from that fundamental simple thing, you can now extend it to much more complicated models, and then that's what this model is trying to do in these spaces. So we might also have sometimes that we don't have that label. We don't have those colours. All we just have is dots everywhere. But we can look at those dots and then try to look at the shape and say, is there something we can read into those dots on the surface? And yes, this is what's called unsupervised learning, where you cluster. Here, we're trying to look for three clusters, and what we're doing is just changing those discrimination functions until the clusters seem like they're self-contained properly. So this is called unsupervised learning. Remember, the first one was supervised because you're giving it the input, where the dot is, and the output, the colour of the dot. This one, you didn't give it any colours. You just have where the dots are, but you can learn some patterns in the data. So these are two important points because we're going to use them as we build on when it comes to language. Why do we do machine learning? Without machine learning, you would have to come up with a recipe to find these things out. So in the example of a dog, you would say, if there's an image of a dog, that it would have some parts of it having pixels that add up to an ear. There might be a snout

somewhere. There will be feet. And only if all of these things are within relative distance to each other, that is a dog. That you would have to write down that program, that recipe, to find that in your thing. But with machine learning, you just feed it the data. You give it lots and lots of examples of the things that you want it to learn, and it learns that pattern itself. And that's why we're going through these major changes in machine learning because now we can feed more data, we have more computational power, and the models can become bigger and bigger and bigger and bigger, and we're using them for many, many different things. That's a major part of this thing. So hopefully then with some of that foundation laid, we can then start thinking about where language fits in. So here we're going to go through natural language processing. We know these through some of the breakthroughs we've been using for decades. Whether for some, the first time was using a search engine, of typing something and getting an indication of where to find something on the internet, or a spam filter in your email system. What would have happened? It processed the text that was in the email and then said that this doesn't look, this is just spam, and it should just go there. Something happened there. In these spaces, there's now virtual assistants, you have chatbots, more and more use, and then translation systems as an example that are on there. We tend to think about natural language processing in terms of tasks. So here's an example. You have an input like [foreign language 01:30:56] output being greetings or elders, and this is translation as a task. You might have, one of the worst days of the year, the output being negative. This then is called opinion mining or sentiment analysis. So you can see, just like when we started and we looked at supervised learning, we're always doing input-output. Something gets fed in, and there must be something that we want the machine to learn to put as an output. Who's the president of South Africa? Cyril Ramaphosa. Just for context for people who might

watch this a few years from now, this is 2024. Before the elections in South Africa. This is question answering and comprehension. You have received the credit of whatever. Output being suspicious. This might be spam detection. And here you get a paragraph, and you have an output as being a sentence. This is something like summarisation. So these are all these different tasks that before the advent of these large language models, these would have been, you would build specific machine learning models for each of them. But with large language models, you can kind of now do all of this almost in one model, and we're building towards that of how that happens. So to do that, just to get people comfortable that machines can process language, let's actually build something together that's very small. So how do we get machines to process language? What we need is these things called features, things that you can feed into the zeros and ones inside. So one is that you can treat words as atomic features. So let's say we have a new language, and this language, just for ease, we're gonna just use English words here, but you could use anything else. We have a new language that only has five words in it. Car, house, three, drove, past. Those are the words that are there. So one way is to use this thing called a vector. It's just a representation that allows you mathematically to express something. So we're going to express that the car is a vector where you have one followed by four zeros. And hopefully you will get the pattern soon. House is zero, one, followed by three zeros. Tree is zero, zero, one, zero, zero, just like that, drove, and past. So just with a vector that is size five, you can represent all five words, and now it's numbers as opposed to dealing with the words inside there. So now you could think and say, okay, how do I represent the sentence? One of the ways that you can do it is you can just say, oh, I can sum or add all of these vectors together, depending on what words show up. So for car, drove, past, tree, house, it just becomes one, one, one, one, one, one, one, because each of those

words show up inside the sentence. If a word shows up twice, you can put a two. If a word shows up three, you'll put a three in the place where it's up because you're just adding them together. The one thing to notice is we lose order. This one list representation, it doesn't tell you the sequence of the words. It just tells you that that word is in the sentence. So you could have lots of different orders inside. But it's actually quite, it's still a useful representation, you will see in the next slide. You can try to deal with the order, and one of the things you can say is concatenate. So instead of adding, you just append, you add to the thing, and now it just becomes this one, zero, zero, zero, just like that. You add it to each other, but you'll see, there's this thing we say there's no free lunch. With this representation, it gives you order, but it becomes bigger. So to now write the sentence, you have an increase in the length of your representation. So that is one thing. What can we do with the first representation where we lose the order? It's something we call term frequency. Remember I said if something shows up twice, you write two in there, right? So because it's a frequency, you can think about document retrieval. So you have, let's say you have many documents. We call it a corpus. You have a search term queue. It's made up of these words, the thing that you're searching for, and you want to return documents that are relevant to queue. What you do is you count how many times the symbols you're looking for show up in every document, and you return the documents that have the highest sum of those words or symbols. So do you see how the first representation allows us to do document retrieval? Because all you need to do is say, oh, I know in the vector in this position, it tells me about this word. This is the word that I put in the query on my document retrieval or search engine. I'm looking for the word cup. And now it just says, oh, of all the documents I'm looking at, which one has the highest count of cup? And I return that to you. So these are the basic building blocks to a

search engine. And what the corpus would be would be what we call an index of the search engine. There are different things that you will find inside here. So what we call this is a bag of words. We don't care about the order. It's almost like we just take the document, we shake it around, we take words, and we just count how many words we see for each of the documents. A feature does not need to be a single word. It can be pairs, it can be triplets of these words, because maybe you might be looking for a phrase instead of just one word, those types of things. So there's ways to do this. Another thing is that in all languages, you have words like the, of, a, that are not really information as per your query. Those are not things that you want to focus on because they show up in all of the documents a lot. So what you might want to do, so those you might want to deemphasize them. And you have this thing called term frequency, inverse document frequency, where you change it, the equation, just to then multiply by, if a word shows up a lot across all the documents, it should just go close to zero. Ja, so I was saying you can de-emphasize words that occur a lot, because they're not really the information. So one way is just to remove them by multiplying them just by the inverse, like one over how many times that word shows up, and then it becomes worthless. So yeah, we've built a search engine, and now we want to get closer to this language modelling thing, right, of what's going on. So in language modelling, what we're trying to do is also understand language semantics. So instead of just the part of we have the document being represented, and we can do something with the document, we want to learn something about the language. And what I'll show in a minute, so is ultimately it becomes math, and I just want nobody, you don't need to be scared. In this case, it becomes this math that you want to estimate this quantity here. This is called the probability of seeing a word, given that you've seen a sequence of words before. This actually has a common name that we all know, and we likely use a

number, like many, many times every day. This is autocomplete. What are the chances, given that you've seen five words before, of every other word that is supposed to come next, and then we just choose the one that has the highest chance, and then we put it there. That's autocomplete that you see when you're typing in all of these services. So this becomes a very interesting thing, because if you can estimate this quantity, you have to learn things about the language, some grammar, the likelihood of seeing some things, so it's a very interesting thing that you want to learn, and this can be used for generating text, which is the thing we're trying to get to on our journey here. One of the things we can do, which has been the thing that I was scared about, but we will do it, is we can actually build one together here. We're going to build a language model live, so don't worry here, this is a Python thing, but it's okay, we will go through this. So here we start, we start with a couple of sentences. I just wanted to make sure that this is big enough for people to see. This is a beautiful day, it is a beautiful day for ice cream. This is going to be our input data into the model we are about to build. So we'll get through that. Don't worry about that, but I'll talk you through. So what we're going to do now is go through these two sentences and try to take triplets of words, so three words in sequence, and move a window across. You will see why. So we take those, and here's how they come out. We have none, none, this. Why is it none, none, this? At the beginning here, there was no word before the word ‘this’. So what we do is something we call padding. So we add to the beginning a token that doesn't mean anything, it just says we were beginning, but because we had nothing before, we're just going to say it's a none. And then another none, and then there was this. Then there's none, this is, this is a, like is a beautiful, a beautiful day. So these are, we're now changing this into different data points. If you've got that, you can now look at count and say how many times did I see the word none and none followed by

the word this one time, given our kind of small data. I can then do this for all of our things. Now you make it into a pair plus the next word. So now you see this is a, we saw it once. Is a followed by beautiful? We saw it once. Is a followed by good, once. Oh, something interesting to note there. Is and a has two times that it shows up in our data, and with those two times, the next word is different. It's not the same word. So we can change this count into calculating chance of probability, right? So if we transform this into probability, this is what's going to come in here. Is and a, given our small two sentences, has a 50% chance of being followed by beautiful, and is and a has another 50% chance by being followed by the word good, right? This is something that's important for us because now we capture that and we know this from the data. One of the things that we can then do, I'll skip that, is we can then ask the machine and say, okay, we're going to start with none and none. Can you tell me what the next word is supposed to be? After you figure out that word, keep on doing this, unroll, and keep on predicting into the future. So we've built the model. The model just keeps track of those probabilities. Everywhere else it puts zeros because it hasn't seen anything. It just puts numbers where it's seen the changes. And once it's done that, so for example, it says, this is a beautiful day. It is a beautiful day for ice cream. If we run it again, it might say, this is a beautiful day. It is a beautiful day. It doesn't say again, it is a beautiful day for ice cream. Why? Because at the is a, it has to flip a coin. One time it will say good, and then the other time it might say, like beautiful. You see, so now you see, we went from just the simple representation that we've used. Hopefully we're coming back. I'm almost done with this one. We're not going to do something like this again. Oh, there's a corner there that's missing. Other things come in. I can work with it if it's a, it's a thing. Yeah, work with it. So it's a beautiful day for ice cream. So if we keep on running this over and over, because there's that branching that happens, it'll

always, when it gets to is and a, have a choice to make. Right, so here's again, is a good day for ice cream. It doesn't go and say, it is a beautiful day again. So that's an important thing because we've learned that and now we can generate new text, almost like new text. Sure, this data is not a lot. What we can do here, we've got some data that comes from government communication information services and we're going to load a big part of their data. It's the Vuk'uzenzele publication and we're going to focus on the English for now and we can do this across this 2 300 articles and we can then do exactly the same thing we just did with those two sentences with that and welcome to live demos. Hopefully there's no error again. Okay, great. And we can do the same thing and now we can start again. We're going to start with none and none and ask it, given now it's gone through all these government documents, ask it to generate, I think generate a text that says there, the ombud will intervene in matters of the outcomes of the past. If I run it again, it will then start black women owned businesses in the churches, synagogues, mosques and all those things. So now we've just taken a huge data set and we've learned this big, what we call transition probability matrix and now we can now start generating. There's no, it's not just learning the math or the statistics. We haven't done any big modelling as yet, but you can still do generation, right? This one, you can't ask it questions and that's where we're going to be going in the next part of like, how do we get to a point where we see these machine learning models that now you can ask a question and then they respond, right? But you can see the building blocks have already started in this space. I'm just trying to see. I've loaded on my side. I'll talk through it as I fix that. So, okay. So we've dealt with now building what we call a statistical language model and now what we want to do is actually there's opportunities for us to change this into that, what's this, into that machine learning problem, right? That classification problem where you're going to have an input and

you're going to ask the machine to learn the output. So one of the ways is that we can start with that process of just say, you have a set of words that came before. I'm going to give you those words and we know now we can represent them as numbers and we can put those numbers into our machine and now we say just predict what the next number is supposed to be and you build a big machine learning model to do this and you run it across lots and lots of different, I mean lots of data and then it learns what that sequence is supposed to be.

CHAIRPERSON: Shall we take a few minutes break and just fix the technical side, I think. Let's do that. Then it's easy. All right, apologies everyone for the technical delay. We're back up and running. So please carry on, Vukosi. Thank you.

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: So yeah, so as I said, we've now, we looked at building something that's simple that allows us to do this autocomplete, but you can turn it into a machine learning problem where you take as input like a sentence, you remove a word and you get the machine to learn what is the word that you've removed. So this is where these models called transformers became very pivotal in 2017 when they came along because they could do this very well and inside them they could then capture some of the things we think like you know, as grammar, as all these things that come from. It's a statistical learning. It learns what are the statistics of seeing a sequence and if you remove a word, can it then predict what that word is and that's where the machine learning part becomes important, right? So with these transformers, it allows us to scale this and then be able to also use parallelised programming to be able to do this very well. So what happens with this is that you have to do something called pre-training. So before we get to something like a Chat GPT, a Gemini, a Cloud or whatever, you need to first pre-train with as much data as possible to learn this, it's called this task of filling in these missing words. So

the first model was BERT that came from Google and with BERT, you would take a corpus, a set of lots of books and Wikipedia in English, train this model and now you've got this model that knows something about language, right? It's not everything but it knows something about language. Then you take that and then you fine tune it. You continue training it towards your final task. One of these final tasks could be spam classification, right? That's one of the things that could be there. Or one is that you can just tune it to start helping you finish your sentences in your email. And this is actually, if you think about when BERT was released in 2017, just after that, some of these features started showing up in people's emails, right? So that's one thing. So now the supervised training or this fine tuning is important because then you can then change it to many, many other tasks. And the tasks that I think a lot of people are interested in is how we get to these things of question answering, right? So now you pretrain it on a large corpus to learn something about language and then you fine tune it or do supervised learning or supervised fine tuning on a task. And this task now could be something like question answering. So here's a data set, for example, it's called the Stanford question answering data set. It has 100,000 question answer pairs. You're trying to get the machine to now anticipate that it's going to be given as input, a question and it must generate an answer that is related to that question. And if you give it enough of these questions and answers or prompts, with probable answers, it learns to connect the patterns again. Remember, the patterns of how do you answer0-ppppp a question. There's things like Trivia QA. This one has 95,000 question pairs. And here's another one, which is Alpaca, which is a recreation of a model where you see you give it, here's an instruction that's here and then give three tips for staying healthy. And then this is the output that you would have wanted it to generate. So the machine learns from being given lots of these examples. And you can create them.

And over time, then the more of these you have, plus the pre-training, the model starts generating these answers for you. So we've gone from just doing spam classification to now doing prompts. And if you do prompts, remember I had that task list at the beginning, you can just make those tasks just prompts. And now you can have one language model being able to be asked all of these different tasks, like please translate from this to this and then it starts actually doing, learning how to do that as part of the data. The thing that happened though that really brought us to most of where we are now is that it was great to get the machine to start, or the models to start generating answers. But sometimes the answers wouldn't meet the preferences of humans. So you might generate things that are shocking, that might not be palatable, that might be biased. So one thing that happened was this instruct GPT work that also came out of OpenAI in mid-2022, where it said that no, what you can do is now, remember you pre-train on a large amount of data, or the missing word filling. Then you fine tune on prompt question answering. And then you do one more thing which is this called instruction tuning, where a human submits a prompt, they get a response, and then they have to grade if the response is good or not. And then this then closes the loop of getting the preferences that humans might have. So you could have these companies having lots of people who are just sending prompts, seeing responses, and then they grade or rate how good those responses are, and then the machine now learns to adjust. If you think about the early days of your big chat GPTs, there was this question of, oh, is it left or right leaning? And it seemed that at the beginning, it might have, from political affiliation, was leaning one way, and as later, as people started complaining, it started moving another way. All of that is because people are trying to adjust this instruction tuning. It's part of this thing called reinforcement learning with human feedback. You reinforce things so that it doesn't do things that are off the rails.

So you want to put some rails that it goes in, and this is also very important, and you see it in some of these models that they're generating. They'll either say, like, you know, I can't really answer that in the way that you want. So you needed to have all of those pieces available kind of to you. There are many language models it is not just the ones that you think from the big tech companies there are many of them and because people are building them for very different things they're one that are completely open source, you can see the data that was used you can see how the model actually gets trained you can see how it gets fine tuned, all those things to other ones that are closed where you can't see all of these things but they're big. So GPT4, the numbers there mean how many parameters you have to train or tweak right, when you have a Y is equal to MX + C which is this linear you have two parameters that you're tweaking that you have to learn the M and the C and then here, with GPT4 you are at 1.76 trillion parameters that must be tuned and you can think about how much training and how much compute, how much data is used inside there. And there's Gemini Ultra and there's more coming and there's other ones that might be small, extra small, like you know medium, large, those types of things. And this has really been a boom that has happened since about 2017 in building these models and getting to the point. At the beginning we might have looked at Wikipedia. As I said with the initial transformer paper for BERT that came out of Google. It was just Wikipedia and some books. Now you're trying to figure out how much data you can get from the internet to train these models. So in this section I'll go through things to consider, for consideration. One is obviously we have to think what's the source of this information that's being used to train these models. Is it credited, to whom, and did those people consent? They know that this might be used in this way. In some ways there's question of is there compensation for some of this content especially if you're thinking not the

like you know right to research or fair use spaces. Is there that part and is also that if we're just using the internet as a training ground to get data isn't it there's also the worst of the internet that comes into these systems, that's another piece on them. The other one for people to note is that you know there's this part of thinking that these are all knowing machines. They are not, they skip also a big part of what we call information seeking meaning that we typically, you get some sources, you read the sources and then you form your own opinion or conclusion. So sometimes people think that these machines are, oh they've done that. And that's not necessarily what they've done. They've looked at through the data that they've been fed, what are the things that are more going to be statistically popular you know, more likely to be a good response as opposed to being what, if you yourself, given a specific context, if you read the source, you might come up with a different conclusion, right. So that's another thing when it comes to these to these things. Then there are with this work that covered and saying like can you make these things too big given that they have blind spots of things they can't see and these might then like you know have challenges with safety of using these models. There was also how we try to make them look like they are human in the way that they chat back and they respond and this might then have unintended consequences and we must think about like, you know, these limits. The question then comes of where is the data that comes into machine learning in general? And a lot of it doesn't represent places like us where we're on the African continent so this is just a visualisation that shows for a specific time. We're looking at data sets that are used in machine learning research in general not just in in language and you see that there's very little that is coming out of the African continent. And this could be very dangerous in that now they're trying to provide context. In building these models if you don't have data that's you could have responses that are not okay. And the other thing

is we can think about this thing of, oh it's AGI, they can chat about anything that I want, but there's all these other things that we don't see. I'll speak to that one so we can move on. There's all these things that are at the bottom of the iceberg just for context or you know you see just the tree at the top and you don't see that there's a whole root network under there to be more contextual to being on the African continent and not talking about icebergs. But you don't see that there's all these other things that go behind the scenes of there might be challenges of workers. We have to feed this Information. Who checks and make sure that there's content moderation? How do you filter the data so that you can clean it? All these things are not things that necessarily the public is thinking about when they use the system right, but this work has to be done. Somebody has to has to do it. To close off, in some ways, as I said, we need a lot of pre-training data and we need to think about where we get it. So we started off with Wikipedia. If we think about Wikipedia from the South African context. I'll give two parts as linking about 20, I think 2019. That's the way we had this table. So this was just thinking about Wikipedia in terms of how many articles are written in South African language at that part. And, if for English, there's six million part plus Wikipedia articles that are written in English, right. If you think about Afrikaans, I think it's already gone over a hundred thousand, but you already a factor of magnitude down. Then you go to Sepedi and isiZulu. You're now in the tens of thousands. And my wife told me this, to remove and put a zero, isiNdebele, for example, doesn't have a local language Wikipedia yet. Right, so if you are using Wikipedia already for the South African languages, you're going to have challenges with representing our languages inside. And remember, for English to, like you know, the thing that you're getting good, like you know, a service is when you're asking things in English and get responses, you have six million plus. For the other languages, Tshivenda, you’re at like 367 articles.

And our demographics, as per the 2022 census, have also changed. Isizulu, it’s almost 25% of the country. It's being Zulu, then it's Xhosa and Sepedi 10%. But the services, for these services, they're not as good as that for English in one part. So Wikipedia is not big enough also. That's to also highlight this. And this is where crawlers come in right? So crawlers is that you have these systems online that are normally used for search engines that used to go, they go around, they go to websites. They follow URLs. So what they would do is they’d follow, go to a URL, find other URLs on the page, but then also download that page and put it into an index or a database. And then keep on following the URLs. So it's expired as they move around and they crawl across the internet and they download this, the data that's on the content that is on the web page, into their databases because this then becomes important for when you then go to a search engine and you type in a query and they find the result and then they show you that oh, here's the thing that's connected to your query and this is the URL that it comes from. So that's where the search engine started, that's what…But now you can use crawlers for different things. One, archiving the internet. So saving the content that's on the internet now. Believe it or not, websites do die, content does disappear and also things change and you might want to then have an archive of what the internet is like at all times. And you have projects that are doing this like the way back machine or the internet archive that are always crawling and trying to keep copies of the internet. Then you can also use the same data to help AI learn. If you can then extract as much of this content that comes from the internet, you can now give these and train these models as well because you now have so much content that is also in some ways diverse because it comes from a lot of content of the internet. You can then use this and you can filter. And this is where services like common crawl come in. So common crawl is an open project that aims to crawl the internet and make that

data available for researchers to use, right. And there's another one called Open Web Text. There's Book Cop, book data sets that just have books. Some of them are using only books that are already out of copyright. Some of them you might find have books that are still subject to copyright. You have social media crawlers as people who are backing up things like, or getting Twitter, data, Reddit, all these services. I think Facebook, a long time ago, tried to, like has been working to close off access to their services so you couldn't actually use API's. And then X or Twitter, in the last year or so, have also reduced, or almost removed completely, their research programs where you could actually get data from them, but that's, that was a big thing. And also post the big, large language model. Yeah, that was 2023. A lot of these services have closed down free access to their content, their user-generated content because they see it as an asset that now can be resold, right? So, you have access to now this content that you can sell to all these AI companies that want to teach their machines things right? So now, for us as researchers, we no longer have access to do research but if you have enough money you can get some sort of access, minus if there's now local policies that say there is a right to research and the platforms must give access which South Africa at the moment doesn't have that kind of guarantee. And you might have scientific paper data sets like archive and PubMed. So common crawl, just to provide context, as I said, it's huge. It's about 250 billion pages over 17 years that have been backed up. And it's always, the common crawl, it's always going and crawling the data set. So even if you might, for example today, block access to your website, but weren't blocking access to your website five years ago, they probably have your data from five years ago and going before, right. So that's what common crawl has. As they say, they're adding three to five billion new pages every month. And for context, talked about Wikipedia, common crawl is 9.5 petabytes. It's very big. And the models we're

talking about when we build the LLMs are actually smaller than this. So what then tends to happen, just to highlight something else that happens is this bias thing, is that you are compressing what you are getting on the internet into these smaller models and with that, that's why sometimes they also still have these blind spots. Because if you think about an image, when you take an image and you make it smaller, it becomes blurrier because you're doing compression, so you lose information. The same thing is happening when you take these models. They are big, they take a lot of energy to train, but they're still not, it's not a one-to-one copy of what's happening on the internet. So that's another thing that brings up bias and say that if you don't have enough of a representation of your language, of your context inside there it just becomes even more stereotypical, right, because of this compression. That's a high likelihood of that. So the models are trained up to a date so you'll see some of these models when you ask a question. They'll say I'm only trained up to X date. What happens to incorporate recent data. And there's one way, there's many ways to try and resolve this, but one is called this retrieval augmented generation. So you now, given everything that we've learned today, you combine a search engine and a large language model. So when you make a query, one query goes to a search engine to say please bring relevant web pages connected to this query. Take those web pages, add them to the prompt that I made, and then send that all to the large language model and the large language model responds faster. So this then requires that your crawler for the search engine still has to keep on working. Even if you're not retraining the large language because it takes a lot of time and energy, but you might, you are able to still respond contextually faster with new web pages because you are keeping your database expanding for your search engine, right. So this is called retrieval augmented generation and it connects us back to some place where we started thinking about

web crawlers. And thinking about search engines at the beginning. Then, in terms of the impact of paywalls and crawling of information. At the like, you know, like in the advent of search engines, there was this thing about having these crawlers and of saying most websites have a robots .txt file and this instructs these crawlers if they are allowed or not allowed to crawl that website or can only crawl certain parts of it and not everything. And it's commonly used to do this and in some spaces it's almost like a, you know, an agreement between all these organisations and in a more distributed way that we do this. But you can have a paywall on top of that, that just is like you know, in this case it's a protection measure, that then say like you know puts an electronic barrier that you can't see my content because you would have to have authentication to see it. And it can stop both a user and a crawler from seeing it because it can just be saying it doesn't discriminate. As long as you get to the URL, as soon as you want to get access to the content, it says no you must first login in one way or another. And this can also stop crawlers that disregard the robot .txt, right, because the paywall can then say like no, even if we don't have you on our do not crawl list, you know, we had the paywall stops, maybe the people in there. But content owners or publishers can get into special arrangements with specific organisations that have these crawlers to say, if we know that it's your crawler, because we can check where it's coming from in terms of IP, we can give it full access to the data without having to authenticate. And the people who write the crawlers might ask for a tag and say, you must tell us where your paywall data starts, or your content starts and where it ends, so that if somebody is trying to, we can use the full information to help us index to make it better lends in some ways to search that content, but if we need to then give people any information about where we got or how we got your page, we only show the content that is not paywalled as part of responding to that person. So

again, it requires that everybody keeps to their end of the bargain if that's the agreement that you're going to. So paywalls is not necessarily that they, in some cases they might not apply to some organisations that are crawling or some search engines, they could have agreements with some of the publishers to say, we get a special access so that we can give you a proper way of indexing your information. We then promise, if that's the agreement, that we won't show the paywall the information to third parties when we respond to their queries. That's one other thing. What does diverse data give to LLMs? It's better at processing things like more languages because you can have diverse input languages, if you can increase that diversity. It improves the context of questions. It's more knowledgeable as it gets a wider range of subject matter, and it includes more background because now you know more about other spaces, not just that it comes from like you know Wikipedia and there's cultures. And it somewhat deals with some of the challenges of bias. So I suppose you can't remove it completely because the internet, even with like you know, if you've got all of it, it doesn't represent everybody as well. So there's some bias that you can reduce but you don't remove bias completely because you have diverse data inside them. Local data, there is a role to play for local news content local language data because it can enhance some of the capabilities of these models. Remember here we're making almost an assumption that everybody interacts with the world in English and that's not true even in South Africa. And the data that has been crawled, as I tried to allude to earlier, if one didn't have a blocker for crawling the data like a year ago, then some crawlers already have this going back 17 years or even more. And even if you put a blocker now, unless now you go to that organisation and ask them to remove your content, it's still there. And with some of these data sets, because they are open and reshareable, you can't really remove your content because it's copied and it's

everywhere, right, in those spaces. But it's important to think about this, that having access to the content also gives us kind of this little, like you know, one of the things that common crawl does enable is a right to research and fair use research, the thing, because if you didn't have that, you wouldn't be able in any real way to say I'm doing an internet study and I have a place where I can just go and search for a word and find all of the content and do that analysis. So there is some parts for research that is publicly beneficial, that these crawlers have a role to play. We work in doing work for African languages and finding content that is written in our languages is important for us and you need to have play, like you know, opportunity to do that. At the same time, private data from platforms is becoming, or it was already a big asset, so if you have content, even if it's user-generated, but that's sitting on your platform, a lot of these platforms are starting to protect it and say like no, they're not really easily accessible to outsiders because it also becomes a competitive advantage. But either you can sell it or you block your competitor from maybe able to train a large language model than yours because you've got users that are using your platform and creating content on there. And for us, again, thinking about a researcher who works on African languages, this has meant sometimes the closing of social media sites for us being able to use, to do research on them, means that we're losing access to people writing or speaking in their mother tongue and us being able to then use that to build new language tools right, because of the advent that now you've got these the new, I guess, internet economy created by these large language models. And there is a lack of funding for local language development and this also then impacts the general availability of tools. I tend to ask how many people, in the first language that you speak at home, have a dictionary on your phone or on your laptop in that language. There's typically very few people. That shows you that we need the development of these tools to even enable

other things that come on later on. So there are challenges in processing these. At the beginning of February 2023, some of the colleagues at Lelepa worked on hey, can we check how well something like chat GPT works on South African languages? So here's one asking chat GPT, during that time, they've improved it obviously, but just showing you that if you don't have diversity and you're not improving your data and your modelling, count from 1 to 10 in Isizulu and the response from chat GPT was, good one, good two, good three, good four. Alright. Another one was translate from Isizulu to English. [Foreign language 02:22:06] and it says, he insults by calling him a coward. And it was actually supposed to be, it's expensive to get you know to fly on an aeroplane. Alright, so and then now sure, these might have been fixed for Isizulu, but just think about the other languages where you're moving towards the margins. That these can still be a problem. Even later when we were looking for some of these things that are odd that doesn't make any sense. Like here, 1 to 10 Isizulu, this is another model, this is an open model, and it says [foreign language 02:22:39] as one and that's not true. So these are things that are still likely, when people find them, yes you might laugh, but then if somebody is using it for information-seeking this is a problem. So as I finish off, this is not just about web data. The infrastructure to train these models now is becoming really the domain of very few around the world. There's even increasingly for nation states, that they can't train these models. It takes a lot of computational power to do it, large amounts of data on them and these language tools are required. You need them because to build up you don't build on nothing, you build on top of other things. So these are important. And there is a need, not just in South Africa but across the African continent, to improve R&D, the local R&D ecosystems, right, in building the people who can do research, developers who can build on top of that research and get it to people's hands, engineers, all those things. Organisations that

are doing local R&D in these spaces and investments to do it. It is not something that is necessarily, if we think about it, that it's going to be done by other people outside the continent because they would want to do it. It has to be something that's built here because of the context of the ways that it makes, that it represents us and the communities that we live in. So yeah, I think with that, hopefully I have been able to take you from you know, fundamentals to how the large language models at the moment work. There obviously will be other small details that are missing. I'm just trying to get people to understand the intuition so that I can inform the panel here and then also the public and then we can go through questions.

MS. PAULA FRAY: Thank you very much, Vukosi. It was very, very interesting. I don't know if I caught everything but certainly caught a lot. I want to take you back to the issue of datasets and bias. I mean, your presentation showed that most of the input datasets are from the global north and I'm wondering a, what is the impact then on bias? But also is this because the datasets are largely based in the global north and that there's a lack of datasets for Africa?

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: Yeah, thanks for the question. So data is multi-level in terms of getting content and you have to deal with publishers and how publishers might share for research. So some might, say no we release some of our content for research purposes only and maybe non-commercial. So how many publishers are doing that? I think there's been maybe some conversations over the last three days about some of the dire state of the publishing industry so that might still be a challenge because people are trying to protect the content that they think, or they know like you know, they can still make some revenue off. The other is just yeah, how easily accessible is it? How many researchers are working on those languages? Because it requires a substantial investment. For me it's

not a cost, it has to be done. South Africa might have language policies, but it is not connected also to investments that work on building up those infrastructures. So that's just thinking about language. The same thing will happen when you're thinking about images that are representative of people here. It's a license, like you know, how easy is it to be licensable? It might be easier for people in the global north to make and say we will release under open licenses and it might be harder for somebody here to say that because of just how tight the connection is to revenue. Yeah and then because if those things are not represented then there's a higher likelihood of bias being within the models that are trained of that data.

MS. PAULA FRAY: But if you look at common crawler, is it possible to determine uncommon crawler how much of the content they belongs to media publishers or is it anonymised?

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: If I remember correctly, I stand to be corrected, you should be able to figure out what the URLs, where they crawled because it's, it has to be, I don't think they try to hide where the crawl has been right. So you would be able to, even though it's like you know 9.5 petabytes, you should be able to write a script to identify all the publishers that are captured inside there. And likely when last they were crawled, those types of information.

CHAIRPERSON: Yeah, I think my mind's a bit blown in the moment, but I'll try work through. I just want to go back to the crawl and just so I understand. And so, I mean, as I understand, let's just take the search engine as a starting point. I mean the search engine crawls the web to index and rank and it will look through the information on each website. So what is going in the index and the ranking? Is this just a summary? Is it the URL or what is in that index?

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: It's likely the content that's in there. There can be, just as I said, like we represented a document at the beginning as a vector. There can be an embedding version of it as well. So this this this one where it's just the numbers so that you can do easy retrieval just to say what's similar to the to the query, but you will also still have the content itself saved inside the index as another part of that database. And then it will obviously be the URLs. There's metadata that might be specific to each search engine in terms of what's important for them. And you could do tagging, that's also connected to oh, here's how when you get information that's connected to this specific tag, you then represent it to the user because it has specific information. So you could have something and say like oh, this person is a celebrity that is being searched for and with that there's other information that actually could be linked to the response.

CHAIRPERSON: So is it in some way a copy of the Internet then if you've got the content?

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: Yes. I think you have to, to build an index you need to have the content you know, because then when, especially having to respond very quickly and present it to the user, for a lot of search engines they show a snippet. That snippet has to be in the database. It's not that now, at the time of the search, you then do a request to the original publisher and then ask to get back the snippet. That is one a way to identify that the index has that content.

CHAIRPERSON: And that snippet is, I would understand from searching myself, I mean, is often relevant to my query. So although it's a big page, I'll get the excerpt that is relevant to my query and that is what may make me click through or not?

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: Exactly. Yeah. And with that, is that there might just again be some more processing to just highlight, okay I'm

using morph as a human to highlight which part of the full text actually is most connected to your query and that's why it comes up.

CHAIRPERSON: And then you did say that that at least you can differentiate the crawler. If you had a paywall, you can say well, I'll allow this crawler in. And as I understand, that often is for search because you want your page to be well represented in results and whether it ranks appropriately or not. So often, even where people have paywalls they'll let them search crawler, search engine in?

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: Yes, so that's I think that in terms of paywalls, are an electronic measure to block access. And we could even go via, what's this geo, sorry geo blocking. This happens in some cases. For example, you could be at a university. I studied in the US. You could be at a university, go to a specific news publishers website, but because they can identify that you're coming from a university, that they have an agreement with, then there you don't see the paywall at all just because you're on the campus and you are logged into the [stream? 02:31:36] There's no exchange of information except that your IP address says you are at the university campus, right? So that is a similar way that if you have an agreement with a organisation that has a search engine crawler, they can come into an agreement and say for you, when we know that it's your IP where this crawler is coming from, we will just give you full access to our content.

CHAIRPERSON: But of course, if you have no paywall, then any crawlers coming in or could you keep out some crawlers and not other crawlers if you're going to display to the public? So, you know, like the media that we're dealing with they're gonna display to the public, they're reliant on advertising. So can you stop a crawler?

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: So you could do the robots or txt as a file and then as long as the crawler respects their robots or txt they can

check inside the robots or txt file and if that robot txt file says you, if you are this crawler don't crawl, and you respect that you don't, the other one is that again you can have electronic measures. If you know what the main crawlers are and there's likely database of people who keep track of where all these crawlers come from, you can just say, as soon as we know that this crawler comes from this address then the website becomes inoperable or something. So there are electronic measures that you can likely make, but then you have to then make an investment in keeping like you know, in some ways you probably might start playing a cat-and-mouse game right, because if some crawlers really want to get to your content then they will find ways to spoof themselves or use proxies to get to you.

MS. PAULA FRAY: So robots .txt is as an honour system. I mean if that doesn't respect you can just go ahead?

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: Yeah, the robots .txt is akin to an honour system. It'd be very interesting if there was a case law in South Africa maybe, I haven't seen it as yet. I'd have to talk to my colleague [Dr. Joker? 02:33:34] about this and say like okay, what do I say, like you know what service? Again, I'm not a lawyer so in this case, but yeah from, at the moment it's an honour system, but it can be taken as an instruction that has been given to these crawlers to say you can or cannot.

CHAIRPERSON: But I think, as you said, I mean it's it it's relied not just on the honour but for you to put the instructions in there. I mean in your experience how much of the South African web is savvy to all of this and I mean, you're an AI person, the kind of typical web developers and website developers are that, that help people put up their websites, are they savvy to the full range of crawlers coming from wherever?

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: That I would not be too sure of as like yeah, I'm not necessarily in the web development space, but in just thinking of convenience, which is normally where we start from, if I was thinking is that there's a one part of looking at the benefits of being indexed, right, that for me, if I want people to find my content, being in an index is good especially of the big search engines. So you might want to make it as easy as possible for them to find you. And to the point, I do have a number of personal sites and with some of the search engines you directly also register with them so that they can, to kick-start the crawling right? You actually tell them that this is where I am and please come here and find our content. Yeah so you could, so there's a lot of things where it's a kind of like search engine optimisation that people do and that might then also reduce the likelihood of them being worried about the robots or txt, but that also right now, as this gets more and more into the public consciousness about how content is used and on the other side for training some of these models, that there might be some people who are starting to ask and say oh, what controls do I have?

CHAIRPERSON: And I understand from you, there are many crawlers and you could identify them and they might have their own specifics that you know it's them that's coming because you may want to let them in or not let them in. So I assume like Open AI, they'll have their crawler?

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: They likely will, yeah.

CHAIRPERSON: And there will be some way you can say, well if it's Open AI's crawler, do I let them in, do I not let them in?

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: Yes.

CHAIRPERSON: Because I think we heard from Media24 yesterday that that they also blocking AI crawlers, such as your search engine crawlers, you've got AI crawlers, you

probably have your own crawler, I don't know. I suppose, for me, what's interesting is, I mean is I understand when you bring AI on to a search engine, that they have to use the search index. I think you put it up that you would put it into their model somehow in order to now respond to fresh information and fresh data. So if they're accessing the index and they're not crawling the web when we put in a query, then it's whatever is indexed by the search engine that…and this seems a bit of a blurring because you've got the AI for training models but you've got the search and you got search with AI where it seems if you've given permission to the search engine you've given permission to the AI?

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: Yeah, in some ways. I don't know if there's specific agreements or contracting that might be coming on within the industry as a whole, whether it's a search engines and the search engines that also have AI in them. So that I don't know, but from a technical perspective, yes, you would think that you there is a benefit that if you've got a very large index, training on that index is a positive because the models become better. And then on top of that, then having the this augmentation of taking, first retrieving from the search engine, that's why it's called retrieval augmented generation, to treatment from the search engine and then connecting it with the large language model gives you also better context right. As opposed to just, if you went directly to the large language model it might respond with something that doesn't actually like you know, if you're using something from the 30th of January and now you ask it when is the South Africa election for 2024, it might then have a response that is not satisfactory to the user.

CHAIRPERSON: So, because this is a question I was sort of had coming out of Media24, is you tell me you're blocking AI crawlers. Is that the training model? But you're not blocking the search engine because you want to be indexed properly and

have your content appear, but if the integrated AI on the search has access to what the search engine is crawled then what are you achieving with the AI?

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: So there's, like there's a, like it depends on the setup, but technically you can have a language model that is trained completely differently and still connected with your search index. That is a big part of the current wave that we're having in AI has been the open source movement that has really went on and said, like we can build open source language models. And now you can take one of these open source language models and then connect them to your search engine without having like that open source model, you know where like you know they would say as much information as they could, sometimes they don't, of where they got their data and they might not be connected to your index at all. And that's possible. So I think, there I'm just trying to clarify that you don't necessarily need to, it's just a maybe the question of technicality that's being asked is saying, if I own the index and I built my own language model as well, you know, what stops me from doing both with the same in doing both with the same data? And that might be the agreement that it is between the publishers and the person of saying, or the platform might say, we announce ourselves as well. We can announce that this is our search engine spider and this is our LLM spider. I don't know if that's what they're doing and maybe that that gives people more control. But in terms of if you own both and if you don't tell people, sure like you know how who's to really know until you have to start doing most likely what is this, like investigation of what is the content that some people are trying to retrieve from inside the language model? But you can, the reason the augmented retrieval has become so interesting is that you can build your own index for something, right? It doesn't need to be a general search engine. It can just be for competition law in South Africa. And then you have an LLM that is separately

built, not by you, and then you use this that when people put up queries the LLM responds with your database of competition law as opposed to just giving general information that would come out of the LLM.

CHAIRPERSON: Yeah and I suppose, I mean to what you're telling is, I just try and understand it, I mean I could have trained on historical data, even the common crawl and let's say people have put up paywalls now, but if I integrate it with search I've got the fresh content. So if that that training of the model is just to be good with language and prompts and certain things I then can augment. So even if I'm not crawling the web now for that model, I'm still getting the benefit of being able to respond fresh. There was a former tribunal member who said, you know, in the future AI will write our judgments, but can we trust Google to judge itself? But who knows, maybe our life will be easier the future. So I just want to go to one particular AI model which I came across, which is not really involved in these proceedings just to understand a bit better, and it's called Perplexity which is, I don't know if you're familiar with it?

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: I think I was talking to one of one of my students this morning about it before I came here.

CHAIRPERSON: Oh well there we go. But this seems to be, it's not part of a search engine, but it does crawl the web for content and I suppose what it, it pitches itself, it's designed to look at fresh content and current events. So it's crawler it seems goes out to crawl and index the web quite selectively rather than maybe the whole corpus, but because it's designed to respond on current events with accuracy, trust and timeliness, you know, I think what was interesting, is it starts to talk about prioritising where it goes to look for information because it wants to be accurate, and also because it wants to be timely. So just from your perspective, I mean, so can a crawler now start to differentiate where it goes in order to get information? As you said, there's a lot of

spam out, there's a lot of junk, there's a lot of hate speech, there's a lot of harmful stuff. So can it be selective in that sense and how that how does that work?

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: Yeah. So let's, we can think this through together by, okay so I'm not necessarily going to focus on Perplexity, how it might work, but we can think about this for myself again. We're now jointly in this room designing a new LLM company. And then we think of like, okay we might want to only get content that comes from trusted places, right. So one, you could think of oh, there's government official announcements that we might want to keep and as such we can white label instead of black, meaning we whitelist. Only use these government URLs as a place to stay. Then we might say, oh here's some news, some good journalists that we follow. Here's what their social media accounts is. We index only their accounts. We index only the news websites that we really trust and we can keep on slowly curating the, basically the limits. We have a list of, these are the only URLs you can use and we keep on adding and evaluating them by human. That's one way right. So this thing takes the human computer collaboration going. That's one way to do it. It's a bit slower, it requires more heads in a room and for some people they might take it as it's costly. The other way to do it might be through click-through, right. So that's why you might have some services that ask you to say, was this helpful or not? And now we say, we'll release version 0.1 of our LLM and now start looking at the usage rates. And as we get the usage information we now start looking and say it seems people really are always drawn to these providers of content. And as such, we're going to start prioritising that our spider also keeps the index fresh from those sources. And the ones that people tend to always say either, no this is wasn't a good response or wasn't good for me or people never visit. Tose get deprioritised, that the next time the spider, it's almost like you give it a like , you give it a budget to say, you

have these many hours today to crawl and these are the ones we would like really like you to get to and these are the ones just become at the bottom of your list. So that's a second way we can think this through right? That we can have a technical solution that does what the human curators would have done, but in a way that is necessarily can always be happening as you're getting user information, as you're getting results and seeing how people are using your system, that that gets goes back. And I have a feeling that is like you, that even people like Perplexity are doing both. You have the human curation plus the machines just having some algorithm or recipe to follow in figuring out where or how to prioritise and de-prioritise content.

CHAIRPERSON: And I'll come to what they say they're doing because I love AI chatbots because they tell you things about themselves too. But if they've coded language like you showed us, I mean this may be a dumb question and I may not have followed everything, but are you basing your trust and accuracy on the site or on the content because if the content is just noughts and ones can that tell you anything about the quality accuracy of what's on the site or do you have to select the site first?

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: So there's always going to be a bit of both here because in terms of, okay using that word notoriety hopefully in the right way here, if to be noteworthy right is that you could be random person with a personal blog and you write about the specific topic and you are the best person in the world, that when people visit that small blog, they know that this person is the expert in this area. Right, you can literally be somewhere in the Eastern Cape, in like you know, right near the sea and that's all you do. So the point is now, if the click-through information, it becomes very apparent that if somebody's searching for this, it tends to be that the small blog, not a major publisher, small blog seems to be where people are really trying to get to, you then say oh, this becomes priority. That's the information

that you get necessarily from that clickthrough information. But that click-through information has to be connected to the zeros and ones of oh, what was the search about and how do we now connect it to the clickthrough? Right, it's not general, it's contextual to in the search, when people are searching for this, there seems to be that the people who are looking at this topic are really interested in the writing of this one small block.

CHAIRPERSON: So you almost looking for signals? You know it's there or you're looking what people are doing but you can't analyse the content itself?

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: Yeah, you need to have something that is giving you information right? You can have something else, like now you might say, and say like okay, I am looking for good writing about cars. There, what would have happened before, is that likely you would have had to teach the model how to, what is the definition of good writing? And now with that, you then can go into the index without, I mean into your data bank, into your index without having to go out. But that's you know, it's a narrow, it's a narrower question than just saying we want something that's fresh that's giving us those things. But you can, there are ways to do that, but you would have had to do the pre-work. Do you have definitions of the questions that might be asked?

CHAIRPERSON: Alright, because interestingly when I did ask Perplexity where it got its information for accuracy and reliability and also timeliness, I'm sure, as you can guess, up come the media academic articles, industry reports, but I suppose what also surprised me was also some social media discussions, but maybe as you say, this is not my discussion, it's someone like you who has a little more knowledge, that they could be prioritising and indexing.

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: Yeah. And remember even with that, you can kind of piggyback on other people's prioritisation right? So if you know a social media, again we're doing the thing of like we're building this thing together and we're gonna spend as little money as possible to get this company going. And we could say, oh there's a there's another company and this company always advertises on their social media platform. They advertise who their top experts are and we just keep track of that list. And every time they add a new person, we add that person onto our crawling list as well. So you can do that. But the question then becomes, as I said, like once you do stuff, like if you do only the curation ourselves, it goes back to the question of bias. How do you evaluate for South Africa? Who are the trusted sources for academic content, for the news content? And then once you then scale it to we have 54 African countries, how do you do that? How do you go across languages? Maybe the most useful content in those spaces is not in English, is not in French, is not in Portuguese. Is that indexed as well?

CHAIRPERSON: And just take this morning from Chris, his sites are invisible. So community media, no one's listening to them because then it seems web crawling, and of course crawling the social media as well, but those are getting blocked off because they're owned by the social media companies. That becomes the primary input into a lot of AI and fresh knowledge.

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: And you do have another thing that is connected, I think with that just to highlight is, even with the media companies, we work, I didn't have a slide on the centre for example, we're working on getting off the ground a collaboration with Ndibela which is a small Chitonga language community newspaper based on Polokwane. And they might, they will have their audience and you can get their audiences, I think it has relevance of who they're trying

to get to. And with some of these we see, and the way we see them as researchers sometimes, is you see these online communities that are very active right, in their language, speaking, engaging, and all those things. And you know, with proper responsible research, like oh it would be great to just observe the use of language, but now that's again gone. And it's still connected with the publisher, that the publisher created the space using a platform, but they also themselves likely won't get the benefit of curating that community that could also help them in understanding how other people are seeing those engagements.

CHAIRPERSON: And this what's interesting. I think, you know, over the last few days it has been highlighted that Google is not so great for vernacular language, at least for newspapers and community papers in South Africa, but the social media sites are where it's all happening because language is maybe easier because you're just typing in. It's not trying to index, assess and rank and determine what comes up. So a lot of the vernacular content and data sets are going to be on social media.

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: I think on one is that there might be different, I think it's great having conversations between, but there also are different maybe goals that people might have versus a search engine, even what a search engine company might be doing in those spaces because in some ways you could think that the technical solutions are universal, some parts of them are not, but how to switch languages is becoming better. We like you know, this we do like just on general research. It's not ideal, there still has to be more work that has to be done in these areas. But yes again, it's the comfortability that you have communities, the things that have been amazing. It's finding communities of people who just write short stories in a language and they're putting it on some social media site, alright and then they're just saying, oh we share short stories for us to keep our languages going, right? And

then we write these things for each other. It's not necessarily they're writing it for the LLM or for us, but it's something interesting that kind of they're doing. And yes, how those things are findable becomes tough in some ways. And this is some of the reasons why having local context is it's important because then you have people who work on the research, on the development, who are in those communities and they know like oh, here's how you can find those kind of things. I can interact with community newspapers in the languages. I'm interested in some of my other colleagues can tackle communities because in the languages that they are, it is not necessarily that a very large crawler will understand some of those nuances.

CHAIRPERSON: And just help me understand this sort of language crawl aspect because I mean part of it, based on what I understood from your presentation, is just being able to do the natural language processing in that language, which is just having a lot of text in that language that you can feed through the machine and it can build a better understanding. So would some of the content, just to bring it back to our inquiry, I mean on vernacular language publishers, Isolezwe or community newspapers just have that value of just being able to build a good AI machine with various language abilities.

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: So there's different things that happen. So like I think for people to understand is that yes, at the moment I think we're limiting ourselves to text just because it's the easier thing to do, but you can do audio, you can do video, you can move into other forms of content creation. But if now you're trying to build a database or a data set, there's different ways you can do it. So one, again you can always do the human curation. People who know where all of the isiZulu major content places come from. Then the other one is then the people who create that content can also tag their content with language identifiers to say, when

you come to this website, we're telling you that the language that is used on this website or in this page is X, there's intentional codes that you can put inside there. Then metadata is useful in some ways. Then another one is that you can deploy language identification. So language identification is now you have other natural language processing models that try to identify, that is the content in this language or that language. So this brings us to filtering. So I might just be, and say for me at the moment in time, I only care about English. And as such, even if I'm crawling the whole internet, I'm always filtering that I check the content, if in the content my language identifier says this is 80% English, that's good. Once I get to 60% English, I don't even care what the rest is, I am just throwing it out. It is not going into my index. And there's a lot of work in building the LLMs to either do language filtering, filtering bad words that might lead to abuse of content, those types of things, of saying let's find these words, let's check in the content that we have, remove all of that content because we don't want to it to poison the model right? And there might be other ways of saying, yes we're going to call the internet and now try to make sure we build enough indexes that are in different languages. As I said, there's a question that comes from either the economics of saying what is it that we're trying to do as an organisation? And again I'm hypothetically, I'm not saying anybody's doing in this specific way. And for us, you know, 50% of our customers are really just interested in the English content. If now will I invest in now creating a specific index and understanding the nuances for a language that is in terms of use on my website number 2001? Like a very practical question and that answer, a lot of times, will likely be no. So as such, your experience might be really very different to when you are you're working with one that like you know in a language that is very highly resourced in those spaces. And as I said, it's not so, now we've identified language ID has to be there, so now do we have tools for language ID for

our languages that are very high quality? That becomes another question. Do we have word lists for filtering for abusive content for all of our languages that are good, right? And the reason I ask this is that likely, as we're going into election period or you would have seen over the last year with some of the social media sites, there's a lot of things that end up on the sites that you go in and say why isn't this content moderated? And it's because the content moderation typically also follows the same type of thing of you do content moderation likely in the languages that you know you have as good, like good off-the-shelf tools or tools you can build very quickly. For things that you think are more the fringes, then you don't have so you can switch languages. So if you want to spread spam or misinformation or whatever, you can then, and for some platforms, just start using a language that is not supported in the automated content moderation tools of that service. And this is a question that, I mean, saying that should bring us to think about how things like election misinformation is likely going to spread around South Africa in the next two to three months. It's that the content moderation tools that have to do with our languages are not as much.

CHAIRPERSON: Yeah that is a sobering thought. I mean one of those, so I mean what I found interesting as well is educating here. I mean as you say, we all got to know about AI in a big way when Chat GPT launched, but it's all the things that AI has been doing up till now which is quite specialised, because you know, I know there's, that Google's C4 data set is coming to frame because there is some litigation now and Courtney showed us yesterday that new sites make up half the top ten, but that is a subset of data of the bigger common crawl. And I can't remember what the other two C's stand for, but something like that. But I mean also just doing my own research, you know, there's again a C4 data set that's there for grammatical error correction. You've talked about the predictive, you know next word.

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: There's another one called MC4 which is multilingual where it tries to go inside C4 but then now filter by language so that you know that here's Swahili, here's isiZulu, part of C4.

CHAIRPERSON: Alright, because I was going to go there in fact and translation is going to be a big application. And I suppose translation, you need to understand the language, both of them, so you need that language capability, but then you also need to know how to translate. So what are the sort of data sets you're using to do that? Do you need something that's already translated or are there other ways?

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: Yes. So for everybody to understand, like to know, so one of the things that's been great in the last maybe five to ten years, has been building new tools for also having representation of languages that is cross-lingual, so not just multilingual but cross-lingual, meaning you have tools actually that you can give two sentences. You can give two sentences one is Tshivenda, one is Xhosa and these services can tell you the probability that those, that's the same sentence just in two different languages. So if you can do that, right, you can then start building up different what we call parallel data sets. So parallel aligned data sets is to build a model that can help you, again let's think about the input output. So you can change a translation model. It's is just a system where you give it input in one language and you want it to learn the output to be in another language. Right, so we're still sticking to the machine learning, supervised learning framework. So what you need then to train that type of machine is as much data that goes from that one language to another. So you need people to either give you these parallel data sets that have all these sentences that are translated, or segments or paragraphs to do that. In the past this would only be people literally have to create them, right? So there would be people who just say, oh we will take this data in Xhosa, translate it to

English as well and then we align it, right. Then we know exactly what sentences are aligned. Now, more and more, what we do is we can get a big data set that has multiple languages inside it. We find the tags of each of the languages and we run through programs that do that sentence-to-sentence matching. And you then say, I only want to keep the sentences that are 80% seen as being likely the same one. So an example of something we did at the at the university is we knew that the cabinet statements from government are published and they are translated into all of the languages. But now we want to do the alignment. And when we do the alignment, we just take this Xhosa statement, we take the Ndibela statement, and now we could just go through each of the sentences and just check which ones were actually when they did the translations are almost direct. And now we have that data. The same thing with Vuk'uzenzele, which I showed, we take from GCIS, we download Vuk'uzenzele from PDF, we extracted, we get all the sentences and then we also align all of the sentences. And then because we tell people that yeah, this is not perfect, it's not a human didn't go through every sentence and check it, we're using a more machine driven comparison, but now you have data that wasn't there before that is aligned across these languages, right? So there the thing that becomes again important is the original data. Can you get that? So, if you have something like MC4, you can likely do the alignment as well. It's just that it'll be more searching for a needle in a haystack to do that while the, what's this thing? The one for, if you're using cabinet statement because you know it's the same statement, you already know that, you're just aligning sentences in a smaller set. So it is important, like if you are thinking about this is that knowing that oh yes, Isolezwe, and here's all their content in isiZulu, but if we assume that the content that they wrote today is likely about the stories of today, I can go to another publisher who writes in English and check in South Africa, and check their

content and see, can I align the sentences because they are likely going to be talking, it's the same, I've limited to a day instead of searching their whole…

CHAIRPERSON: Right. Yeah that's interesting because they're reporting on the same event so there may be a lot of similarity.

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: Yeah, you will throw away a good percentage because obviously if you're having, it's not that people are trying to translate each other, but some of them might be that it's the same thing that is just being said in two different languages.

CHAIRPERSON: I see we're almost out of time I had maybe my dumbest question for last, which was, is there a difference between web scraping and crawling or is it really the same thing?

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: You have to scrape to crawl, right? So you crawl, you move across that's why we say because you move, as I said you follow the URLs, so it's almost like you're following a web. But to build the index you have to scrape the site because then after you've crawled, you scrape the content then you put it inside your index.

CHAIRPERSON: Alright, at least clear for me now. Paula, please.

MS. PAULA FRAY: So maybe it's an unfair question, but a lot of media obviously concerned about the impact of AI on their sites. Do you see it in the same light?

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: Put me in a bad position. I'm always trying to ask them for access to data. Like we need to have an informed public and we also need to make sure that our values are represented in different ways. And we need we are going through, and this is my trying to answer by not answering directly, we're going through a period where I think we are renegotiating who pays for the internet basically. I think that seems to be the big part here is that

we're going through that process at the moment and it is worthwhile bringing the public into confidence of understanding the economics of how internet services work, what all of the pieces are, so that when choices are made people actually understand why choices are being made. And I think a lot of that had been just hidden as we just expect three services and they just work in some ways and the first time you interact with a paywall you're like, why are they doing this to us? We just had access to this, but that's that is a major thing. So when thinking about AI systems, I think it's more than just simply economic. It's also thinking about, are they fit for use? Are they safe? And do they really represent us in the way that we want to be represented? At the moment it's, there's a feeling that like you know once you explain to people how they work, they start asking questions of like, but what happens when you do X? And those are typically the answers that a lot of people can't answer, right. And that's where having a space where I think researchers, for kind of in the right to research and fair use, can be, you know, given access in a good way to build up these things locally is good in the same way. At the same time, yes, we understand it's your copyright and IP rights and you need, like you know, people need to be able to enforce them in the way that that they can, but it shouldn't be, for me, at the complete disregard for the development of our local languages and that's a, it's a very tightrope to walk. It's a very tight trope to walk and as I'm at the University and I have to say I'm also at a private company that is also building language models, and I live that contradiction.

CHAIRPERSON: I suppose what you're also telling us is there are the tools to differentiate who's coming in for what purpose?

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: You can identify those, it's just that if people, like because you can provide what we call a header, just telling the site oh, I'm accessing you from here. But then there might be things that don't

announce themselves. So I'll talk about scraping, just for things, so you can scrape a website and some people might have technical tools that just say like oh, if the sustained downloading of scraping of our content for a specific amount of time, yes one, it could be a denial of service attack, but another one is that it might be somebody scraping the site so block that person because we don't know them. But there are technical things to do that. We might do it because we're like oh, this is interesting. We're trying to do some work at the University so we just want to have a local copy of that content we're not going to every time do that, but then we know for some services they'll say oh, we saw that you just went through a thousand pages in five minutes, that cannot be a human. It's a machine that is doing something and because you are not one of our known spiders, we will now block where you're coming from, right. Some of them can do it like completely, like say forever, this place doesn't get access, but if you're doing that you then should know on the other side that it might be legitimate and now doing the blocking. So you see how it again brings, you need to have the technical prowess to figure out who you're blocking and who you're letting through.

CHAIRPERSON: And that may not exist with every media company?

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: Yes. And in terms of the level of what that is, it might be different per company because it's yeah. So somebody might be doing it for nefarious reasons, but somebody might be doing for legitimate like you know, reasons of saying oh, yeah this is just for me, I'm doing some work and I just want to have a local copy of all of these things and it's unfortunately we'll go through a thousand thing. And then some, it doesn't mean also all crawlers are bad that you don't know. Some of them you have, just for people, there was, for everybody, there was a service that used to, I think for the UK politicians, just keep track of when they delete tweets. So at the beginning, on face value you don't be like okay, why

would you care about that? But it's very interesting what gets deleted especially from a person who's in the case there would have been on the social media platform as a politician, as a public person, not as a, it's not their private, like you know it's their persona as a politician. And now, what you wanted to do, there was these spiders that just go and keep track of what their politician has written and they just keep on checking again and say you know, oh this person deleted it now, but you can go to a website and see this is what they deleted two days ago. So I'm saying, like you know, they might not announce themselves in the way that you want, but it's not, it doesn't mean they're not doing a service that actually is for benefit.

CHAIRPERSON: And may even be looking for misinformation.

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: Or doing something on building language resources. You know there's other things that are going on that are just not yeah, necessarily what we think. So it's just that, as I say, like it requires that nuance.

CHAIRPERSON: Well, I think we need to bring this session to a close. I'm sure we could go on all day, but Vukosi, I can't thank you enough for the time you've given us and the public and the media. I think, as you say, you know, I think we all need to learn and you've certainly done a fantastic job in simplifying it. And also giving us some of the real complexities of navigating this world. And I think your question was interesting about, is this about who pays for the internet? You know, certainly a lot for us to think about. And we do have some of those players coming next week so I think you've also given us an insight that we can also engage them at a higher level than we would otherwise. So thank you.

MR. VUKOSI MARIVATE – UNIVERSITY OF PRETORIA: Thanks everyone.

CHAIRPERSON: So with that we'll break until two when we have the mail and Guardian joining us for an hour and that's the end of the program for today so if you can join us at two I think it'll be an interesting session with one of the publications that really has led investigative journalism for a long time in this country.

ADJOURNMENT: AFTERNOON PROCEEDINGS ON 6 MARCH 2024

CHAIRPERSON: Welcome back to the afternoon session on day three of the Media and Digital Platforms Market Inquiry. We have with us the Mail and Guardian, I think it's fair to say an institution in South African media and CEO Mr. Thembisa Fakude welcome and welcome with your team as well. Maybe if you can introduce your team and then we had sent you areas that we thought would be useful to discuss. If you could just give some introductions and then go ahead.

MR. THEMBISA FAKUDE-MAIL & GUARDIAN: Okay thanks very much James and Paula. Let me just quickly introduce Simon who sent us an introduction, he is our shareholder and big boss. Then I've got Douglas White who is in charge of subscriptions, head of subscriptions at office and my left-hand side is James Smith who is in charge, he is our chief digital officer and I'm the CEO. So that's that. They are going to be coming in as we move along. It's good that we have Hussein here. Hussein is an institution, memory of institution so we are in good hands in terms of giving clarification and explanation of the Mail and Guardian. So I'll shoot. I'll try not to take too much time because most of our work was done for us in terms of parameters. So we have topics that we have to discuss and we'll stick to that and ensure that we save a bit of time for question and answer

at the end. I think that's key and very important. Just as another word of introduction. The Mail and Guardian I'm told just before we came in here by Hussein that we were the first online publication in South Africa at least news wise and that was in 1990. Africa, not only in South Africa but in Africa and by 1999 we were the leading or the most visited website, news website. But of course this has since been, a lot of new players and kids on the block who have to an extent not only added interesting news publications but also, they've made us to be much more competitive because of their competitive nature, what they bring as competition in the business. Of course we don't regard them as competition. They are colleagues. We are all trying to do one thing which is informing the public as much as we can and we work with each other as close as we can. But of course over the years we started competing for premium resources, as has been the case. So the first topic that we were asked to tackle was the Mail and Guardian's business model and experience in the distribution of news content. So I will take the first question and then of course as I've said my colleagues will help and if I leave anything out, please do come in. The first part was about Mail and Guardian's views on the trend and print of online news distribution including views of the news conception in South Africa and transition to digital consumption. Like all traditional news media platforms we have been facing challenges and those challenges are mainly because of the digital platforms that have come in and most of whom are now tipping into that limited revenue space that we have traditionally tapped on. And I think the biggest challenge that all of us face as news media is that we, we've got two mandates basically. One is to inform the public which is more social responsibility part of our business but at the same time we are also expected to make money and that

is where the challenge is. So you've got the newsroom that's got its own ideals, they are part of this egalitarian society all trying to build. And they're not driven by profit but they're driven by principle and the need to continue keeping our audiences up to date with what's going on around the world and in our country. But importantly, in the absence of a vibrant political opposition in this country, media unfortunately has become or is playing that part. In other words our investigative responsibility has to an extent moved to a point where they are playing a certain type of role in terms of keeping our audiences and the South African public in general up to date on what's going on. Case in point is the Zondo Commissions and other exposés that were or commissions that were as a result of the media's work in terms of exposing certain social political and economic challenges in this country. So given that we find ourselves in a much more difficult situation. You know, one because of the economic dwindling resources but importantly is that whereas in the past we were amongst the very few credible platforms in the country and we will go to platform now we have other people who are doing similar stuff which of course we encourage, we are not saying they must shut down in any way or form but Mail and Guardian finds itself having to swim with new kids on the block who are doing similar constructive work in terms of informing our public and therefore the consumption of news is diversified. Whereas in the past there was very few platforms you could go to, particularly in the way of weak publications now, there is no weak publications but there is minute by minute publications. And the other challenge of course that we face as a weekly publication is that by the time we tell our stories, it's no longer news, because somebody else has told it 17000 times before you tell it on Thursday night so that's one challenge. So I think the only competitive advantage that we

have and we continue to invest in and we try to improve on it is our investigative platforms. That one is not easy to break, because it takes resources, knowledge, context and access. I think that will be my contribution to the first part of the question. We can come back to questions and answers if there are any further contribution if my colleagues want to add anything additional. The second part is going to be tackled by Scott which is the insight on social media revenue streams for niche publications.

MR. SCOTT SMITH-MAIL & GUARDIAN: Good afternoon, everyone. Scott Peter Smith, I'm Chief Digital Officer at the Mail and Guardian. So I've been working in digital media in either management or leadership for a good 12, 13 years or so. A bit longer actually and so I've seen the booms and busts. I've seen the different challenges we've had in terms of news distribution online. We've tried lots of things. We've explored lots of different ways of doing things but the buffer for me was always we had a larger product backing us up so there was a newspaper, there was a broadcaster, one of those scenarios. So the pressure on me was to make money and find revenue streams but it wasn't always, I didn't have a news organisation hanging by those strings at that time. So we're in a situation now where my position is okay how do we finalise the digital transformation of the Mail and Guardian, the print product, we'll get into the different products. The print product won't be there forever, at least not in its current form. So the job was to me, there's not necessarily one digital platform we can work with in order to solve this problem. We've got to deal with, what products do I have available to myself and what potential revenue streams do we have, that can maybe offset the loss of this print product, which has been carrying for the last 40 years essentially. So what do we have, we've got the website, we have newsletters,

I've got, I've got social media not in so much as a product, but an online product, we've got multimedia things we can build. So when I say product what I mean is essentially like a service, or a good or something definite that we create ourselves that we can sell directly. So when I say social media, we'll get into that question now of how we are approaching it and our metrics for success in that area, but essentially these are areas where how can we sell things. How are we selling this news, how are people accessing and that's not a one horse show. That's okay, we've got a web of different things we can deal with in this kind of scenario. The models we've got available to us in that area is models we know relatively well in advertising of course and the second one we're dealing with is subscription which is quite nascent at the Mail and Guardian and then of course we're looking at other partnerships such as the Telcos for example in terms of distribution. But then you start spilling into the area that we're all here about today in terms of how does that work. How does that revenue share work. The situation I've got now is the Mail and Guardian audience, just take the website audience of the Mail and Guardian which is literally 200 times the size of the print audience in terms of the sales and in terms of traffic. That's not including the relatively significant, not as big as some local competitors but relatively significant audiences we have on other platforms that don't belong to us such as the X and Facebook and so forth. But still we're in a space where we've got 200 times the audience, you think that would be a valid revenue model in terms of what we've got available to us to leverage off of. That's clearly not the case and that's largely why we're all here for this kind of commission. How do we tackle this issue. So we're in a space where we're exploring other things but my efforts, the way we can build out the team in order to do things. I'll give you a direct example to the

second question, how we are approaching social media. Until quite recently, social media was largely seen as a referral tool. So for example even now X is our largest referral but nevertheless it's largely meaningless to us in terms of the business or revenue. So if I want to take all the efforts now in a newsroom, it's okay we've got 1.1 million followers on X, that's okay for us, look at the size of the audience. But the engagement there, that kind of clicks we get, the percentage we have, it doesn't do much for us. We cannot rely on that revenue model, so that revenue. What I mean by that revenue model is we rely on people coming back to our website so we can monetise it in some way, whether it's advertising, or whether we're bringing through into the funnel, ultimately make it into subscribers which is still a valid model for us and we're still pursuing it. However my approach to social media is not as a referral tool. I'm not looking for necessarily traffic there and as soon you make that declaration, you're in an interesting space, because you're like okay well, what are these platforms doing for us. And if I'm going to put resources into a social media team and I'm building out a multimedia team and there's lots of money involved with that. If I want to put say you've got a salary cap of R100k a month and great, you're growing your Twitter following, you're growing your Facebook following. You're getting engagement there. People are aware of your brand. There's a dotted line reason to be there. And their audience there does matter. It's not that it doesn't matter. It keeps us valid, it keeps us interesting. We can do things there, but essentially what we're doing is we're starting to create content for those platforms specifically. You find the use case today is there's been quite a drop in referral traffic, that's not just for us, that's for quite a lot of media organisations seeing a drop in referral traffic from these platforms. So what that means is I can even tell

from my own behaviour, it's largely quite rare for me from one of these platforms to click back into the website. But nevertheless you're in a situation where you have to be there. You have to be, you can't not be on social media otherwise you're even more dead. Otherwise, particularly in the model we're exploring, which is a commercial model, we are advertising subscription based if you're lucky enough to be donor models and you don't need an audience. Good luck to you, that might be nice for now but it's probably not going to last you. At the end of the day you do need an audience. So my approach would be, I'm not just looking at social media as referral traffic but I do need to find some other way to offset the cost, offset the building of that team. Same thing that goes for multimedia and I'll give you an example. In my previous employ I was at Arena Holdings, I built the multimedia team there, the video and podcast, we built that up and there was some resistance to it at that time and I said let's just do it and we found a whole new big audience there, hundreds and thousands of people that wanted to watch videos and listen to podcasts. But again to my earlier point I was backed up by primary print and they had some of their broadcast channels there as well so we were backed up by that sort of clout. We had an existing audience of millions that I could leverage off to build that. Eventually when we built that team, I was in a space of five or six, sometimes seven million views of our videos a month on YouTube. And we would be in a space of making about R50,000 of an expenditure just on salaries, we had a large team, Joburg and Cape Town pumping stuff out every day. That's not sustainable. It's sustainable in the bigger picture that is the Sunday Times behind us or something like that but even now that isn't sustainable. That team is no longer there. Same thing with podcasts we were doing well but there's simply no revenue there. So the

same thing with the website. We can be in a space where when I started the Mail and Guardian the programmatic advertising was not set up properly. It's now set up largely properly. It's obviously moving, evolving space but even so the, setting up as well as we can with the kind of audience that the Mail and Guardian is used to and has, we still cannot leverage appropriately that kind of advertising using our Google Ad exchange for example. I don't know how transparent you want me to be on our numbers but it's far below what we're getting on print. Meanwhile we've got again 200 times the audience. So just in terms of answering that second question I could carry on for a while but these are my experiences with huge audiences, huge outputs, lots of views but the model just doesn't support as it exists from the Google infrastructure and the Google ecosystem, doesn't give enough back to us for that to be sustainable for us. That's even taking in consideration the very good programmes they have in terms of helping news organisations. We did get some funding from YouTube to build some kind of content, which again these are very useful programmes and tools but you cannot build a viable news business off of that. I'll leave it there for now.

MR. THEMBISA FAKUDE-MAIL & GUARDIAN: Thanks Scott. I'll give over to Sam to deal with the third part of the first topic which is the Mail and Guardian's view on independent investigative journalism in South Africa, challenges, costs and digital sustainability. But before I do that I think what Scott has said is very important and I think we all fall into the trap of the pressures of being part of the hip crowd by joining the social media but our audiences are actually not really active on various social media platforms and that's something I think we never had a discussion on Scott inside the office in terms of which particular platforms on social media should we concentrate on. Because we insist on being the Mail

and Guardian, we insist we're going to continue targeting certain audiences. But those audiences are my age and we are more present on Facebook than Instagram. So shouldn't we also be gearing our strategy towards that if we want to maintain that audience but of course the plan is to expand and look at other audiences. As we do that, I think we need to be very cognisant of the fact that we still are the Mail and Guardian which is consumed by a certain type of an audience and we shouldn't be too ambitious going to Instagram and TikTok and other platforms where we don't have that much presence as a business. So Hussein you want to take that part of it.

HUSSEIN ALI-MAIL & GUARDIAN: Thanks everyone. So investigative journalism goes to the heart of what the Mail and Guardian is all about. Many, many years, we'll turn 40 next year and it is well known that many of the country's foremost investigative journalists as well as editors have emerged from having had experience at the Mail and Guardian somewhere in their careers, whether it's someone like Ferial Haffajee who was an intern, to [unclear] Makhanya who was with us to many of the other sort of editors who are now editors of some of the mainstream publications had the experience at the Mail and Guardian. A younger person I guess is someone like Raeesa Pather who also I guess a few years ago did an internship at the Mail and Guardian. But I thought I'd just go back a bit and just run through a sort of quick chronology of how the Mail and Guardian now finds itself where it is today going back to when we launched in 1985 as you all well know by Anton Harber and Irwin Manoim after he left the Rand Daily Mail, and set up this sort of publication telling the stories that no one else wanted to tell. And it very quickly built itself up with a reputation of a publication of good high-quality journalism which I think is what has kept us within the sort of mix of

mainstream publications out there. We're small, we always believe that, where we actually barked larger than our bite, we are, we feel that we're a part of the mainstream news agenda every week. Then in 1990 as was pointed out earlier, we were first news publication to go online in the African continent that was led by Irwin Manoim back at the time. And quickly we developed a very strong digital audience online before anybody else could. Between 1990 and 1992 there was this relationship with the Guardian newspaper from London where they started putting the weekly Guardian as an insert into what was known then as the Weekly Mail. We started out as the Weekly Mail and during that period we became the Weekly Mail and Guardian because it carried the insert of the Guardian newspaper as an insert in the print edition of the Weekly Mail as it was then called. In 1992 that relationship formalised and the Guardian became a shareholder in the publication and that was when the name changed to the Mail and Guardian. At that time as always is the case in our organisation there was a shortfall of some cash and we sold 65% of our online presence to what was then known as MWEB, Naspers effectively. And from 1992 or thereabouts until 2006, we operated where News24 owned 65% or call it Naspers owned 65% of the online version of the Mail and Guardian. From the 1992 onwards already one could feel that the Guardian's interest in funding independent journalism in this country was beginning to sort of dwindle as obviously they started experiencing their own issues back in London with this whole move from print to digital. They were beginning to feel it back then already and in 2003 they decided to sell to a local shareholding, they wanted to localise the shareholding in the Mail and Guardian that's when [unclear] from Zimbabwe came in and acquired the majority shareholding in Mail and Guardian. He effectively inherited this organisation

where we owned the print edition 100% but the online presence was owned partially, majority by basically our competitor Naspers who at that point was launching News24 and it became a very difficult relationship to manage between 2002 and 2006. One could hardly send an email because anything you did online needed to be vetted by the competitor. You couldn't do anything online. I think 2006 the MDIF assisted us and, that's the Media Development Investment Fund assisted us and we reacquired that 65% stake from Naspers and eventually we owned the online presence ourselves. I would say from the period 2006 to 2012 were the boom years for print media in so far as the relationship with its digital output and presence were concerned. We were doing fantastically in terms of our print circulation. I think we hit the 55,000 copy sale a week, sale. Douglas would sort of be quite envious about now, given where we are now. I think we had Ferial Haffajee and Nick Doors as our editors at the time. But also online I think we got to something like 300,000 users, unique viewers online but more importantly we got something like over a million rand a month in digital advertising revenue which was significant at the time which was very good and we felt very bullish about the future of publishing online and the whole shift from print to digital at the time. I think the fact that the M and G was one of the first publication to go online back in 1990 also gave us the view that we always needed to be ahead of the curve. At some points along this timeline I think we overinvested in our digital presence. I mean the acquisition from Naspers didn't come cheap so it was quite a significant investment. Thereafter we always wanted to stay ahead of everyone else. We invested significantly in our online presence. Online luminaries like Matthew Buckland for instance led this whole charge. Chris Roper led this whole charge in being able to publish or publish

more online. I recall very specifically Chris Roper saying you want to get this to 2 million unique viewers, we'll get it there. At that time we got to about a mill and when we wanted to get to 2 mill, 1.8 viewers, we got there. And things seemed to be working out very well. Then about 2013 we noticed this decline in advertising revenue for online and we couldn't figure out where it was going. Because we were serving the same significant page impressions, number of ads on the site but the money just wasn't coming through like it used to and this is when all these programmatic different platforms started to launch. There was a sense that, and Google sort of started investing a lot more in their technology stats that started attracting revenue that ought to have come to a publication like the Mail and Guardian and it just sort of took most of it away and left us very little. But this decline in online revenue commenced in about 2012. In 2013 we couldn't invest as much as we used to in our investigative journalism team and that led to the launch of Amabhungane. You all know Amabhungane started by Stefaans Brummer and Sam Sole has become a formidable investigative hub in this country. We're fortunate to attract donor funding to be able to support their project. The M and G at the time contributed quite a lot to the entire platform. I mean I always enjoyed the fact that I was a founder member of Amabhungane, founder trustee, which I think was fairly significant in the South African media sector and I think it's played a significant role. Regrettably because we weren't able to generate the kinds of advertising revenues we could online and we were beginning to see the shift in terms of print revenue and copy sell decline and print advertising revenue decline we could no longer sustain the co investment in Amabhungane and we had to withdraw. I think there was also pressure from donor funders to not have a trust associated with a commercial entity, which is

effectively what the Mail and Guardian was and we had to then acknowledge that there had to be a clearer separation from the commercial intent and then Amabhungane launched on their own, continued on their own and attracted more donor funding. I think the concept of donor funded journalism was really Amabhungane and that's where the whole, that's morphed into a lot more now. The only other thing I wanted to add is that the M and G has always been at the forefront of establishing these kinds of journalism platforms. It really has been by default, because really, we wouldn't want to disassociate ourselves with Amabhungane. We launched a health journalism platform called Bhekisisa. That happened in 2017, 2016 I think. And again we, due to the fact that we could no longer sustain or afford journalism as far as health care was concerned, we then had to separate. Recently you all know that the M and G is part of a trust called the Adamela Trust which launched The Continent and The Continent is a digital PDF publication that goes out every Friday. Also funded. So experience at the Mail and Guardian will probably be written about in an MBA or a thesis somewhere in future in terms of how not, sort of through our own natural desire to want to be the best news publication out there, but because of the lack or the decline in advertising revenue that we experienced in print, that we experienced, notwithstanding that we invested significantly in our digital platforms. There was just no revenue to be associated with it, to a point right now where the M and G literally struggles along. The fact that we can put out a weekly, a solid good weekly newspaper and a good online presence is testimony to the team that we have back at the Mail and Guardian but in terms of the business model, to fund this has been completely decimated by the likes of what Google has extracted from earning value of the content that we produce through our cost, that we

invested in over many, many years, that we continue to invest in because of the role we played in the South African media sector or the South African political sector and environment and what we've contributed to journalism over the years. It's sad to find us all in this position and we hope that by intervention of what we're experiencing here today it will make a difference to bring back to the M and G publications what is fairly theirs in terms of its advertising area.

MR. THEMBISA FAKUDE-MAIL & GUARDIAN: Thanks Hussein. Doug, do you want to go on competition for ad revenue, subscription revenue and user data please.

DOUGLAS WHITE-MAIL & GUARDIAN: Thanks Thembisa. Hi, my name is Douglas. I'm the head of subscription and circulation at the Mail and Guardian and unlike my colleagues who are media people I'm basically a salesman. That's what I do and in order to do that one has to understand one's audience. One has to understand how the audience works and I think that it's necessary when looking at the Mail and Guardian and its transformation since this time that we look back at our historical model, specifically starting at the print area and moving towards the digital aspect of it. One of the things that I think is important is that we have to acknowledge that the print media itself is in decline. Fewer and fewer copies of every publication except maybe Farmers Weekly are being sold on a basis and that means that the audience continues to shrink. Now that of course was exacerbated come 2020. I have the figures that can show you at the start of the year we were sitting at one particular circulation level. By the end of that year it was a quarter of that circulation. We have been able to pull stuff back but it made a significant dent in the funding capability of the Mail and Guardian, so what I'm going to focus on specifically is the attention engagement of users on

the digital platform. It is very different from how the printed audience engages with their newspaper. When you buy a newspaper, you've got it, it sits on the table, you pick it up, you read it several times a day and then you put it down, maybe somebody else does. When it's a digital medium, you're looking for it directly. It's in your hand, maybe you share it, maybe you don't but it's not an easily distributable product. When you've got people coming to the Mail and Guardian you need to make as much use of them as you can. When it comes to advertising, print vs online with print it's all about how many copies you sell as your circulation figures, it's where its coming from. With online advertising it is how many hits you're getting sure but it's also who's getting hit. Right. I mean and these massive digital platforms have algorithms such that they can tell you what kind of toothpaste you're more likely to buy, just by your searches and what other products you associate with. Because of that it's forcing other organisations like the Mail and Guardian to start trying to make their own inroads into discovering that information, into getting that data. Obviously, there are platforms that are available. We use a free service called Google Analytics which allows us to get that data. Some of it is our own sought, some of it we get given to us by people who are engaging with the Mail and Guardian. But because we, the cost of developing that sort of technology we will never be able to keep up with a digital platform, digital organisation whose entire business that is. The other thing that we have to look at is specially how people online engage with content. As Hussein mentioned back in the 1990, we developed this website, come 99 we were the biggest website in South Africa. Website, not just news. Website. And that's because at the time people were engaging with media in a certain way. You move forward 20 years and the culture and the behaviour has changed partly

because of the way these websites are being packaged and presented to us. I'm sure everybody here has been on social media. It's a stream and you sit there for hours and you stroke and stroke. It doesn't mean you're going to be clicking through or anything. Most people are just happy. They think they've got enough information from the 15 words above an article and the headline, and they scroll on, they give their opinion. They don't care where it's coming from, they don't care who's sharing it. They have access to it, that's what they see. I view it like a one arm bandit. You pull, slots machine goes, hopefully you get something out of, maybe you don't but it's a habit-forming situation so the way people engage is once they're on a platform they don't move beyond it and this has had an effect on news organisations from a vast aspect of things. News24, Mail and Guardian, pretty much any news organisation that exists, cannot compete there, and because people aren't moving, they're not going anywhere. So of course with the declining audience, what we've as a result seen is a declining audience in print, and declining audience online. Not of people accessing online but of people consuming independent news online just because of the way that they engage with online activity. Once again, we can look at costs and figures, 2018 we were getting a certain amount form print and online. Today we're getting a third of the revenue from our print. Everywhere we're having our funds squeezed and then look at the type of audience. One of the advantages of having print in the space, sure you had your subscribers that would get their copy delivered to their house every week but you also had the retail outlets out there so while somebody was walking past down the bread aisle maybe they saw an article or a cover they thought was interesting and they'd grab it. That was always the majority of the sale. You never had more subscribers than you had retail outlets. That is what

this new model forces us into so we're trying to make subscribers out of people who would ideally want to be once off purchasers and that's a challenge that I think needs to be looked at in terms of how we move forward. As Scott mentioned a reader revenue model is vital for us to continue existing because with declining revenue from advertising, we need support directly from the people who are consuming our product and how do we do that. We introduce a registration wall, right that as I was saying to get the information we require, that gives us names, that gives us email addresses. Maybe sometimes they give us their company name or what their interests are. And then you've got a paywall model. Now the paywall model is effectively, you block people who don't pay. You don't get access to the news if you don't fork out of your pocket and this further limits the dissemination of quality news that's available to the public. So I'm just about to wrap up but in closing I think one of the important things as Hussein did mention was that in the early 2000s our digital advertising revenue was in the millions. It is now less than 5% of our advertising revenue. That on the back of a declining print which is also seeing declining revenue and on top of that the cost of digital advertising vs print advertising is sitting at 5%. You might sell a full page in a print publication for R100,000 some will come to you it's the exact same size, the exact same dimensions, it's just being viewed on a phone, and people think no well I'm not going to pay you more than R5000 for that because it's cheaper elsewhere, and it's massive, it's a difficult situation for news publications to be in. Thank you.

MR. THEMBISA FAKUDE-MAIL & GUARDIAN: One other thing we need to add is that of these interventions which were [audio dropping] organisation. They also crumbling. Paywalls, subscriptions so now you have platforms that are actually

giving information, valuable information for free. So we try to protect ourselves and come up with these interventions which we would have hoped there was going to be some sort of solidarity within the fraternity to ensure that we continue in this manner but they have been dropped and people are looking at other creative streams so we are a very vulnerable business right now. Very essential service, we have to continue providing services to the public but the sustainability models that we have, financial sustainability models that we have really needs some overhaul and I think that will take the entireness and hence the importance of us addressing the challenge presented to us by Google as an industry, because by and large the reason why we find ourselves in this situation is because of such big giant media organisations who don't want to be called media institutions even though they are media institutions, coming in and basically exploiting taking what we work so hard for and fund for free. So it's key that we join forces. I'm glad it's not only South Africa that is doing this, the world is doing it. The nice thing of course is that Google seem to be listening. Notwithstanding of course they're not willing to pay as much as we would like them to pay, but I think we are towards, a step towards the right direction in terms of them listening and understanding our plight. Hopefully, that will change. I'll give over to Scott to deal with the advertising and user engagement.

MR. SCOTT SMITH-MAIL & GUARDIAN: So there's, obviously in this kind of environment there's a few ideas we discuss in the newsroom and say okay what's your roadmap here, how do we put these kinds of things together. Part of the roadmap is there some cost cutting measures to be involved, not necessarily staffing but in terms of say for example the print product for example and that idea was tabled in the sense, okay say for example we don't print for every

second week, for example so we're saving a big chunk of money which does a lot for us but again that's not a model. Right. Cost saving is not a business model. But interestingly in that discussion, you know editorial guys were like, it doesn't affect me that much. Their story is still going to get published and they know that their audience in their particular stories are still bigger than print but of course there's a certain loss there for them, their workflows for example are still very print centric so there are still some adjustments there. From my particular side on the product side I was like more investment in terms of what you want to do so I was happy with that but nevertheless I'm still very cognizant of the revenue replacement what we have available to us. Of course the commercial guys in that meeting were very concerned in the sense that they know that the loss of that print product, valid or not so in the marketplace in terms from advertisers is far more valuable than the digital product. There are certain things that we're looking at replacing that and we're talking a closer look at our PDF version of the newspaper. The digital edition, the paper as such and yes there are certain things that we can build that increases user engagement there. It's a better reading experience, there are certain things we can add to that experience, like saving articles for later, downloading offline. There is certain editions we can do there. And also, we can provide stats to the ads that are in that paper in a much more meaningful way to the advertiser. So there are still some little things we can discuss and it's still viable for us and we are moving in that direction and with the other products that I mentioned earlier, there is a road map there, there is an uptake. The subscription is showing some sort of uptick force mainly because we invested in our tech stack, we invested in our onboarding journey, we invested in our database management, so these kind of mechanics behind those models

are working far better than they were a short time ago. However in terms of the revenue replacement, the runway we have in order to do that, is quite short and we're in a space where things are declining probably more rapidly than anyone here thought. So again what do I have immediately available to me, I have my website traffic. Soon as you're looking at your website traffic and you're starting competing in the local space for that. You're like okay there are certain things we can invest in as the tech stack issues there, there's the SEO technicalities, we can build around that. There's some publishing SEO we can do better as well, some headlines. So there's things that we've improved and are of course working on those things but how does that translate. My problem with Google and the aggregation that they have ongoing is yes, we want to be there, we want to work with Google, we have to. But in my view if we were in a space, we can dig into analytics, for example, we were roughly around about 5% click through. 5 – 6%, it fluctuates a little bit. But it's quite low compared to the kind of audience generation that a platform like Google sees from the Mail and Guardian, sees from the kind of content we produce. We work quite hard to make sure that our stories do appear where they need to appear. And there's work involved. Never mind the big, long editorial journey to actually produce the stories that we produce. So my view would be if we are in that kind of space, it's not a viable business model for us. We can't rely on money from Google. There are of course all kinds of conversations in place in terms of maybe Google shouldn't have the snippet for the story. People do want to know more about the story, they do have to click through in the sense that they're not reading enough of what they want to read by the search result snippet. Again we can't control that. I have to be a little strong in my wording, like aggregation without compensation is almost like

a land grab really. It's almost we're using our hard work in order to build our audiences and build our revenue streams. We're giving you a piece of that pie but it's relatively small. It's not in a sense going to be able to support what we currently do. So again we're in a space where we've got opaque analytics or at least opaque processes for the most part. The communications in our regard are not good and you can look at some other platforms like Meta for example who did, who made some very drastic changes in terms of their approach to news recently. Mail and Guardian was not in the space where we had built our business on Facebook but I know some other organisations that did and of course that affected them greatly. I would make the argument that of course, Meta doesn't necessarily have an obligation in terms of keeping publications going but nevertheless if there is a balance to be made in terms of the value they create out of our hard work and what they get out of it. So in a competitive landscape from South Africa because we're an English publication, we're competing with every other English publication, most of them are English and of course we're competing with international publications, and in the news space. In the western world most of those are English. How do we do that. Mail and Guardian is not in a volume game, we're never going to be in a scenario where we're going to be starting to see user traffic in the 30 million or 40 million things like that whereby models like Programmatic start making a little bit more sense. But we're never going to be in that space. We're not going for that space. It's not the nature of the Mail and Guardian's journalism to be in that space. I think we will essentially be in a space where I'm building other products. We are building newsletter products or multimedia products whereby we control a little bit more in terms of the sponsorship of said products. It comes down to a direct play with our readers.

We've got a small readership but they're committed, they're dedicated. So okay how can we leverage that and that's the space we're in at the moment in terms of our competitors. We cannot compete with the News24s of the world. We're not a breaking news organisation in that space and we can use an example say for the, we covered the Senzo trial for example which did very well for us. We've got a lot of views on that story but it doesn't translate into great revenue, nor did it translate into a great conversion into subscription. So yes, we can cover the stories and some stories in our coverage we can compete but there's no scenario we're going to be pushing out the amount of volume that News24 does for example. We have to start asking ourselves what we are building on. How are we going to make this work, subscription is part of that, building new products is another part of that. Whether we're going to be able to do that in time before we have to necessarily replace the print revenue, we'll find out soon enough.

MR. THEMBISA FAKUDE-MAIL & GUARDIAN: Before I give over to Hussein to conclude on the sustainability of investigative journalism, I will answer two questions. The first one is the competition between Mail and Guardian and the news organisations and the competition between Mail and Guardian and the following news media. I don't think there's competition. I think competition suggests that there's some kind of hostile existence between ourselves and other news organizations. Because we're all clear in terms of our mandate and what we're supposed to be doing and why we exist. So that is clear, my only challenge is not necessarily the competition but is the centralisation of funding. Which creates a problem. Because if a, for me it's another way of a back door entry of monopoly and we see that happening by the way. You know. A lot of people mean well in terms of funding but the fact that we're going to have certain very

big funding organisations who then in turn will directly and indirectly expect certain type of coverage is problematic. I will rather see a diverse or diversified funding model or presence but at the moment what we see more and more is we are seeing these big organisations, funding organisations dominating the business, particularly in Africa. And as a result you are starting to see and hear a certain type of coverage in terms of news and this is what Mail and Guardian should and is trying very hard to stay away from, that we need to look at independent funders, but importantly I think we also need to start Hussein changing the way we think in terms of how we sustain our business. This whole idea of continually approaching funders to sustain our business for me is problematic. I think we must revert back to the old organogram of news business where you had the managing editors who were supposed to do business including investing in business so that they can sustain the business of news instead of us sitting and expecting that we will go and fundraise. I'm of the view for example that we need to start businesses that will sustain the business of news so we have to start using actual scientist for example to tell us how we use our money. So most of us go up, we go out and fundraise, we have enough money but wait for it to finish then go again and fundraise. Plenty of times, authorities have done in the past not of course withstanding that they cry with us when we all cry but they're quite well off compared to us in terms of you know, financial existence. So this is the model I think that will work. Because this funding model I think as I said, it's got a potential of the political domination creeping into media business and we see that happening as we speak. There are certain organisations that are dominating the funding space which I think is problematic. The other point is about our competition with international or foreign

media organisations. We have no choice. The fact that we're on the World Wide Web, that in itself that we want to play with the big boys. We need to prepare ourselves with the eventuality and the challenges that come with being in the World Wide Web. Because World Wide Web in itself is a competitive space, and you cannot but compete with big media organisations. But again, it's not about competition, as far as I am concerned, it's about providing alternative voices so that we amplify journalism and we amplify the mission that we've all signed up for. What I think is key and very important though is that whilst we are competing, whilst we are competing in inverted commas, we need to realise that there are also synergies. And I think this one thing that most of us as we continue to compete have forgotten about that news media is about synergies and that solidarity within journalism helps subsidise certain costs. You don't have cover the story at all costs even though you don't have resources. We've seen the partnerships that have emerged over the past couple of years. Panama Papers, a case in point where the media institutions around the world came together and they decided on how they will split the spoils again in inverted commas so you cover what's relevant to the audiences etc so these are the kind of things that I think we need to continue doing. I was speaking to some of our colleagues in South Africa that we don't have to skop each other all the time. There are certain times where you don't have to skop me. You can give me the story. I'll give you the story next time and I think that culture of sharing resources, sharing stories in this business is key. Sometimes you might have funding coming your way which for example might be specific looking at other forms of storytelling. I must be able to say I don't deserve it but the other guy deserves it. Again this is the culture that I think is key. The fact that we've got institutions such as SANEF, the

foreign correspondence association of Southern Africa, these are the platforms that are supposed to be facilitating that, but instead I think they are dealing with other issues which again are important issues, but again one issue which has to be included within these organisations, the umbrella organisation which represents journalism in this country is to look at how they can facilitate sharing resources when they come their way and encouraging that culture within the membership. Finally we have tried at the M and G to ringfence journalism and how did we do that? We have done that by establishing, again thanks to Hussein and team this Adamela Trust the purpose of which is to ringfence journalism and fund journalism. Because we have a lot of people who want to give money but they don't want to give money to an institution and facilitate Thembisa's sports car but they will rather want to, you know fund journalism and what Adamela is doing is going out to fund journalism so that ringfences journalism. So even if the institutions do collapse for whatever reason but journalism will always survive and I think that for me that's key and another way of how we can protect what we do best, and separate, and minimise the risk that has befallen some of our colleagues where when the business collapses everything else shuts down and I think what we have started at the Mail and Guardian is a new way or method of ensuring that journalism survives and if it does succeed we are likely to continue doing what we're doing best without hinderance. Hussein, do you want to take the last one, sustainability?

HUSSEIN ALI-MAIL & GUARDIAN: Thank you, so the views on the risk of corporate and government funded investigative journalism. I don't think we've seen many. We have had instances where corporates have asked us to or sponsored a certain project or a certain objective and we've always had very

strong editorial people in the organisation and it's led to lots of debate and exchanges between a potential sponsor or funder and the editor themselves because we've got a very specific idea about what the output ought to be. So in terms of the risk of corporate investigative journalism I think that you can build in place, mechanisms to protect the independence of the journalism and the independence of the output. But I think what concerns us more is government interventions and I'm not sure if any of you know that recently government has once again raised the issue of it may be in disguise but the issue of a media appeals tribunal was mooted may years ago. I think it was back in 2014 when the, and when the press ombudsman's office established the inquiry to determine the BEE status of the news sector and at the same time we also came up with a very specific code that the establishment has written and that ensured the protection of, it revised the code in so far as the press ombudsman is concerned. But I think that we got to a point where we were able to maintain an independent sort of self-governing institution like the press ombudsman's office which works very well. I think it continues to do superb work until today. I think what government has been trying to moot is an establishment of a media charter. And the publishers have obviously taken a very strong stance against it. We engaged in the discussions with them. But if ever it does come into being we've insisted that it be a voluntary charter so those types of interventions do worry us and we would resist as publishers as much as we possibly can any form of government prescription or regulation that will expect or envisage media to toe a very specific line. I think we all know that press freedom is enshrined in our constitution and we will do everything in our power to ensure that remains in place. Our views on the existential threats to news media sits there. We've heard a very excellent

presentation earlier today about AI. You know and how that works and this whole question about crawling and scraping and what they do with it with the content that, takes a lot of investment on our part to produce. One is going to have to ensure that there are rules in place in future to protect the sort of IP ownership rights of content, what that content, what happens to that content, what gets done with it but this is a changing world. This is how the ecosystem will change and who knows how this will look ten years from now but certainly in so far as this inquiry is concerned and what Google has been able to get away with over the past ten years that went by unchecked is important for us to now intervene and to ensure that we reclaim and we get the business models right again where we earn our revenue based on the traffic, based on the content, based on the investment that we put into our journalism. We reap the yields of that investment in a fair and equitable manner. Right now we believe that it's not fair and equitable at all.

MR. THEMBISA FAKUDE-MAIL & GUARDIAN: Thank you, that's our presentation. Thanks very much for the opportunity.

CHAIRPERSON: Thanks so much. I think you've raised a huge amount. The Mail and Guardian is an institution and you did well Hussein. I think it's been a breeding ground for investigative journalism and quality editors for the entire ecosystem. Just on, I suppose for some people, we've still got some investigative journalism happening, doesn't it look fairly healthy. We had the Gupta files, we've had, I mean what has been lost in a sense of investigative and just to give an example Paula was mentioning to me how there used to be the consumer investigative work that was done in the 90s or late 90s. I don't want to give away your age. But Bhekisisa's depth in health rather than surface work. Just looking

back at the Mail and Guardian and what it did do and as you put it rightly, it was doing stories others didn't do. I mean what do you think has been lost in this period as the squeeze on the media has occurred.

HUSSEIN ALI-MAIL & GUARDIAN: I think the control has been lost in that these projects and the really good platforms are funded externally. Bhekisisa will know, [unclear] was funded by the Bill and Melinda Gates foundation, and inevitably whilst you try and protect it from any intervention there are obviously going to be mandates that need to be fulfilled so what you've lost is the ability to publish free without fear, without intervention, without the sort of having to report, and I think that that whole ability to be able to just investigate free of any interventions has been lost because one inevitably has to go back and adhere to whether budget line items were there, adhere to project output, one has to adhere to them and the same I guess, and the fact that none of these projects are ready to fund from South Africa. They're all funded internationally and that brings about a whole lot of challenges, and also brings about the opportunity as opportunists would take advantage of is to claim that these agendas have been driven by the west as they would say. So it does take away the ability to be able to comfortably proclaim that this is free and independent journalism which is I think what we had in the past and which is what a real commercial viable organisation would be able to proclaim.

MR. THEMBISA FAKUDE-MAIL & GUARDIAN: Just to add two things. One is once we strive to be successful in investigative journalism but two things have come to the fore over the past couple of years at least in South Africa. One is the threat that whistleblowers face, many of whom are killed these days if they whistle blow and that's the main source of news or breaking news and if that is

in a way destroyed you are therefore likely to continue doing what you do best. Secondly is in South Africa again you have the political parties and intra political party point scoring or settling of scores and many of, because of the generalisation of the newsroom you find many journalists falling into that trap, indirectly or unwittingly pushing or promoting a certain political agenda. And we see that happening all the time and I think that's one thing that, two things that we need to look at as we address the challenges faced by investigative journalism in South Africa. How do we go around that and how do we avoid one, being the facilitator setting of scores within the political parties as you try to do your work but importantly how can we partner with whoever it is to ensure that whistleblowing as a phenomenon is encouraged in this country? We can either tell the story and make sure that we amplify that because I think the law authorities want us to do that, you know, making the public aware one that is a good thing to whistle blow if you see anything wrong but importantly that there are platforms that protect whistleblowers and until that is amplified you are likely to have this very important source of news and information becoming smaller and smaller. Specially, we have a lot of cases which are currently pending which are not concluded nor followed up properly by the legal authorities in terms of who killed who. Then you're going to have a situation where there's going to be discouragement from the public to supply us with what we need for us to make good stories and hold this government accountable.

MS. PAULA FRAY: Can I just take you back very, very quickly, so in your heyday, I don't want to call it your heyday but in 2012ish when you had about 300,000 viewers online and about a million online in advertising, fast forward to today, is that still the same ratio?

HUSSEIN ALI-MAIL & GUARDIAN: Completely the other way around. So you've got over a million people looking at our site, in fact more, because on so many different platforms, but the revenue has shrunk. Because the bulk of the revenue gets extracted by Google and the remnant amount that we get through programmatic interventions is miniscule. Completely miniscule so it's completely reversed. Despite the fact you've got all the inventory, all the traffic, all the ads being served in this inventory, the yield that we get is far lower than what it ought to be. I mean we've done a calculation that had all things been equal then this business would be completely sustainable. Had we attributed growth to revenue numbers we earned back in 2012 in relation to the traffic that we've been able to grow our sites to, had that business model stayed this would have been completely sustainable and South Africa would have been better off not only by virtue of the fact that taxes and everything it earns would be local but the point the quality of journalism and the craft of journalism would have been enhanced. So it has had a real impact on journalism as a sector.

CHAIRPERSON: Thanks Paula and I'll come back there but I just want to pick up one question on, I suppose we heard from another alumni, Khadija the other day and she was sort of also indicating that as resources shrink maybe the focus shrinks geographically as well and in stories that are further away more expensive to report on more difficult, become lost. She also mentioned the SABC closing so many regional bureaus. I mean do you think just looking at the impact again that obviously we may not lose some investigative journalism but we're losing stories from secondary towns, rural areas at least at the depth of her investigation that we now afford some in the metro.

HUSSEIN ALI-MAIL & GUARDIAN: So I mean I shudder to think of the amount

of stories we miss out on especially at local government level for example and quite frankly if we had, anything similar to what we were able to do in the good days that type of investigative focus at local government, it probably would have enhanced local government. It probably wouldn't result in the chaos that one sees happening at local government because a lot of the stories just go unreported. I know that there are a lot of digital platforms that have come about now that try through the use of data, assemble a better story, a better picture of what is happening at local government level. I know specifically there's a platform that looks at qualified audit reports for example at local government and that has again happened by virtue of digital interventions but it doesn't tell you the real story, it doesn't tell you the story about what happens to the individuals who are living in a particular area and are most affected by the fact that they have not had water for five days in a row. It doesn't get to the real story, so yes, absolutely there are a lot of stories that don't get told and to that extent, government enjoys this picture of everything looking quite rosy where in fact the real reality of it and the truth is that things are terrible down at local government level. We just don't get told.

CHAIRPERSON: I like the way you framed that, I suppose we don't know what we're missing. But we're clearly missing and I think we've heard over the last few days as well is where there is that vacuum then misinformation can easily spread so I suppose from that municipality there is only one version and it's that that's going out.

MR. THEMBISA FAKUDE-MAIL & GUARDIAN: I think it's also important to understand that we're not just breaking news but we're also a conduit, So over the past we've had partnerships with various organisations on other platforms

and I think this again should be resuscitated in a way where we partner. So if you know that your story is clear and is likely to change the lives of people you need, we need to partner, at least for those who've got an amplified voice to tell that story and we are lucky to have a new editor in our office who has been publishing a lot of partnered reports and I think it's something that's key especially with resources becoming smaller and smaller. The need for media institutions to partner in terms of covering their stories, is becoming more and more. Television is doing it at the moment if you've noticed, you'll find that the correspondent sitting somewhere in Iraq working for CBC will be covering for Al Jazeera because Al Jazeera hasn't got presence and you just acknowledge that this is a CBC reporter reporting on behalf of. I think it's something that we need to start doing as print organisation where important stories where ordinarily we have the voice can be rehashed or republished at other institutions including the Mail and Guardian.

CHAIRPERSON: I think, I mean you're the media specialist not me but I mean that makes a lot of sense. Is there at some point a loss in diversity of voice if it's only partnerships.

HUSSEIN ALI-MAIL & GUARDIAN: I think there is certainly a risk of the decline in the diversity of voice when you're looking at partnerships but I think it's important to note that the diversity in voice is already happening when you're looking at the algorithms that you're being fed. You pick a news story, you're not going to see another one because you pick that one so we're already heading in that direction.

CHAIRPERSON: I thought it was an interesting point about corporate dominance because as we heard from SANEF we sit with four big media houses, English, Afrikaans, it's not a diversity of language at a national level at least. But

I suppose is corporate dominance an ill. What we do see or at least an observation from those that some people do buy into the media for the influence that the media has and they're willing to fund that, out of their own pockets, so it's just sort of fast forward if things don't change in your world. What might the media industry look like in five, ten years and what do you make of that likelihood?

MR. THEMBISA FAKUDE-MAIL & GUARDIAN: I listen to vernacular news and radio. So I listen to Xhosa FM, it's like permanent dial at home and in my car. They tell stories that you ordinarily wouldn't hear on English mediums and that's where South Africa is all about and that's where South Africa decides. And it's a challenge that I don't know how it's going to be addressed. Often, we get surprised when we see a certain outcome of elections mainly because we thought that what we have been fed from the English and Afrikaans mediums was a reality and it's drummed into our heads so many times that when a different result comes, we start getting surprised. But also there's a danger in that because it could in future question the credibility of electoral system in this country, because we have one dominant voice telling us over and over again like it's been in the States because they all speak English. We've got one dominant voice telling us over and again that the governing party or whatever it is have lost, have lost popularity and as a result they're going to deepen their support and if you do get a different result that could be a problem. So I don't know how we solve that and we have tried in the past but again English dominates and I don't think there is much we can do to change that reality. It's quite a complicated one.

HUSSAIN ALI-MAIL & GUARDIAN: I think we have to acknowledge that the sector

itself doesn't present as investment or corporate at the moment. It simply doesn't and if you look at the publishers, we can understand Naspers has taken a long-term view and they are able to do that because of the reserves that they have accumulated and other investment opportunities that you know is able to help cushion the losses and we know it's losses as well. It's public. Caxton have got the printing presses I guess that are there and that helps them. But in terms of investment opportunity this particular journalism or the sector doesn't present as one opportunity other than the fact that it may be the sort of personal agenda of an owner or seeks to achieve something else by owning a media platform. And it will take a while for it to adjust. Having said that I listened to like Michael Jordaan who says this happens in any industry. It happened when DVDs became obsolete for example and CDs came in. You've got to find a way, find another avenue and he turns around and says to media, find the business model. Go out there and find the other business model that's going to make it a success but in the current way it's formulated it's not there. It's just not going to be there.

CHAIRPERSON: And this is not a DVD.

HUSSEIN ALI-MAIL & GUARDIAN: Just on the point of corporate funding I think it's important, I can't remember, I tried to google the name of the publication but there was a publication at the end of last year that was fully funded by an external corporate. And for some reason he decided that he no longer wanted to fund it and he pulled the funding and they were all out of a job in a week. So I think that there is a very real danger of that sort of model. New Frame.

CHAIRPERSON: I think that's a good point because any funding model even if it's, you know Bill and Melinda Gates focus can change, priorities can change …

HUSSEIN ALI-MAIL & GUARDIAN: The open society example where there's been

the transition from [unclear] to the sun who's now decided to change the entire model of open society across the world and as a result South Africa suffered because there's just no funding available and not going to be in the next year or so until it's reestablished in a way in which he's happy with. So that's, a real-life example of how things change very rapidly when you rely on donor funding for journalism.

CHAIRPERSON: Maybe just one last question on a similar note. What we've heard from quite a lot of media is if we have to unpack the algorithm debate it's more around, I can't rely on this, I can't build a business on this, because much like the funder it is unreliable and it's up and down and I need to invest to build something solid. I mean, just to get your perspectives on that. I mean if there was, if there was a partnership where you feel there was fair value then that's another dimension that's still important.

CHRIS ROPER-MAIL & GUARDIAN: Dying all the time to bring AI into this conversation. So look I mean I'm very pro tech in general given my position and everything and there's obviously different facets of technology so our experience with AI in terms of their different areas there that are really bad for us and there's potentially opportunities there for us as well. The areas that are bad for us that I think about quite a lot actually and in terms of some of the models that we were talking about earlier today. Just looking at traffic for example as a simple example, that I'm sure lots of other media colleagues have brought up in these talks in the sense that the click through that we rely on so much that so many of us talk about in terms of the traffic perspective conceptually may not disappear completely but it's certainly going to diminish. I can even tall that from my own behaviour, in terms of Search I use ChatGPT a little bit more, not necessarily just

as a one because it's easier for me as a consumer if I'm looking for an answer for something, when I say answer quote unquote I just mean it gives me the starting points that I'm looking for. And there's no click through there, there's no model there from us as a publisher. However in the context of the Mail and Guardian for example we're in the fortunate space if you choose to see it as such that we've been around for forty years. We have a very deep long archive of really good journalism in South Africa. There are questions that [unclear] are we protecting that IP or how do we want to play the value of that IP. I'll give you an example. A couple of months ago there was one of our writers that saw one of the stories that they wrote for us, reproduced on another platform but it wasn't stolen. It wasn't a copyright issue per se. It was written by AI. And AI had used us as a source to produce that story and it was very, very similar. Same quotes that were built out for a particular story. And you know this writer came to us, what do I do and I was like let me explore that, that's interesting and it's the first time I'd had to deal with that. There are those kinds of questions as well, are we choosing to make our archives available to the bigger picture. It's almost like 20 years ago you would ask the same questions do we want Google to see us or not. I think we're in the same space now with AI. I would suggest that given the coming proliferation of fake news and generated videos, generated audio brand matters more than ever. In the sense that you want to be in a space where the news organisation such as the Mail and Guardian and others around the world, they're already trusted. We're in a good space in the sense that we've got a long track record of trusted journalism. We can leverage off that with the coming fake news on site which we are going to experience particularly on platforms that we rely so much for in terms of the audience and traffic. We could play it in the sense

that it could be beneficial for us. There are other opportunities there for us if we wanted to start looking at, we've discussed a few times today around the proliferation of English as the main language. There are opportunities there for us to use AI to redistribute our news in numerous other languages, particularly in South Africa. AI, that's the thing, they've cracked the language code for the most part and that's the benefit of AI in the sense that it taps into our natural workflow quite nicely and quite well. So there are opportunities there for us in AI with the deep archives that Mail and Guardian has. So there's two sides of the coin and it's going to dissipate our model in many areas but there's also opportunities there. What that looks like of course I'm not sure yet.

CHAIRPERSON: Thank you very much to the Mail and Guardian team for your time, and your input. And of course your service to the country over the 40 years. That was very interesting and I think insightful and it's important to get the perspectives of different types of media because I think the challenges differ between them. Just for tomorrow we start with the National Community Radio Forum then we have a range of community media, Naledi news, Mpumalanga Mirror, Marula Media, Limpopo Mirror and [unclear] Phakathi. So that is tomorrow's programme and then on Friday we have Arena and the Daily Maverick who will be here as well. So with that we'll close for the day and thanks again to all the stakeholders who came through the course of today for some very interesting proceedings.

END OF PROCEEDINGS ON 6 MARCH 2024