Interview: Participant-9

 **I: What discipline or subject does your research fall under?**

F: Right, ok, difficult to define really. As I said, I think I'm Humanities, I would describe my research as being to do with engagement with the media so it might be media studies, 50% of my research (which I know you don't want to discuss) is to do with history.

**I: No, no, if you want to talk about history you've got the right guy here!**

F: Yeah? Well I have transcripts, thousands and thousands of words of transcripts which I would love to have a tool that helped me with that. Because I'm using the British Library's newspaper archive which means these days, because it's all digitised you can take thousands of words from newspapers and I have a collection of thousands of words

**I: That's almost like historical social media data, isn't it?**

F: Yes, it is! It's something that we've not been able to do before.

**I: And actually physically…**

F: …physically and of course you had a limited amount of time so there was a limited amount of data you could collect and it was easy enough, because you were transcribing it you were reading it at the same time. But now with all these resources digitised and I can use search terms, I'm mainly a qualitative researcher but any sort of content analysis of what are the terms that are coming up or even things like sentiment analysis would be lovely. There are a couple of articles out there, there was something that was published last year going, 'Hey wow, we've got all these digitised resources for history and we've not really figured out how we can use them.' But that's like the Brave New World.

**I: Yes, it is a Brave New World in terms of two disciplines, two kinds of – I don't really want to use the term 'disciplines' because we are all doing interdisciplinary work these days, in terms of computing, scientists are now very excited with the potential to apply computational analysis techniques, particularly with regards to natural language processing, sentiment analysis, machine learning because there's now an abundance of data and there's a significant motivation for how this can be utilised by – I'll use the term 'Social Scientists' but obviously in your case this would be a very interesting thing to do.**

F: I mean there's also the problem of contextualisation because of course you get an article and it says it's from the a certain newspaper from the 23rd of May but you really need to see it within the context of the page to realise. So digitised sources bring their own problems.

**I: Yes, it's not a clear case of now you've got an objective focus on the area you are interested in; you still have to take into account these contextual issues. Would you consider your research with the media under the field of Social Sciences?**

F: Yes, probably. The other half of what I do is very much, has been for about ten years, to do with use of social media to engage in politics

**I: With lots of participants…?**

F: Lots of participants, of course all using anonymous names, some men, some women, some not wanting to out themselves as either. But that's an enormous amount of data, again, I have done some analysis, I've done a conference paper and an article out of it but now I have this enormous dataset sitting there and I'm thinking 'I really want to do something else with that.' And the same goes – I'm working on a collaborative project, you can collect all this data.

**I: It's not the collection that's the problem it's the…**

F: It's the how do we then analyse it. And mostly we're qualitative workers so we want to find out about how women are dealing with it, what their emotions are.

**I: It can't be something very deterministic, it has to be in the content?**

F: Yeah, yeah. And it's more like taking this tsunami of words and being able to..

**I: …manage it.**

F: Yes, yes and to explore it with key words but not just key words. One of the major problems with analysing things on Twitter sarcasm and irony, users are incredibly well educated, they are very, very different from other parenting websites, very, very well educated, they write*)* they are very – they use irony, post-modern. So when one says 'well **of course**, you don't know – written on the page, yes –within the context of the conversation it's obvious but if you are doing some sort of key word search, no, it doesn't come up.

**I: Again, that's very interesting. I'd like to go back to the tools you use and your research. So what role does social media play in your current research activities? I think you've already said quite a lot about that but I wonder if you could talk to me a bit more about how you identify sources of social media, how you capture it…?**

F: Identifying it is really – it's a snowball effect somebody might discuss a blog so I'll go and look at that blog.

**I: Kind of like a viral thing, if you had like a heat map it would be kind of like – there's a lot of activity regarding what I'm focusing on with this area of my research so I'm going to focus on here…am I right?**

F: Yes, but if somebody then says, 'Oh, go and look at this blog page' or something, then I'll go and have a look at that. And then I might start thinking 'ok, this is a useful blog because they are always talking about these issues' so I would follow that as well.

**I: So this is very interesting because the full title of our project is 'Tools to support Social Scientists linking social media data.' So if we used the term 'social media data' for not just the typical almost hijacking of the word 'social network' with things like Facebook because, as we're aware, social network wasn't invented with online social media. So it's very interesting that you – am I right in saying that you almost cross-validate your findings, if, say someone refers to another blog post ?**

F: Yes, , we did a project that keeps coming back and repeating on us about blogging, which was really interesting because we wanted to look at the type of person who blogged aand of course that was very difficult because one of the reasons that people don't blog is that there is so much official disapproval of it so we almost had to go undercover to smoke them out and that was very much searching, we searched a lot of blog search engines and blog tools to find – and then we did the snowballing thing because they would be linking to each other.

**I: Link to a link.**

F: Yeah and then we set up our own blog to try and communicate with them because it was very, very difficult to get them to talk to us because they were so frightened we would blow their cover.

**I: There's the whole ethics and privacy issue too.**

F: Yes, but we found the blog was very helpful because they would come on and leave anonymous comments so we were able to get access to them in that way which was very interesting.

**I: In terms of the traditional, well hardly traditional but things like Facebook, Twitter, LinkedIn, YouTube, would you ever go to these sources of information for, I don't know, looking for a relevant theme. You mentioned women talking about infertility, so would you by any chance look at the Mumsnet group on Facebook and just see whether…?**

F: Yes, we'd also use Twitter and, not me, but one of my collaborators is enormously busy on Twitter so she probably would be doing the Twitter thing but Twitter we would use, Facebook

**I: And how do you capture the social media data that you are working with?**

F: Very, very old-fashioned way of copying and pasting and sticking it into word and things,

**I: A very manual process the**n?

F: Very manual process, yes, I'm sure there are lots of other ways I could be doing it and I don't.

**I: And when you gather this information and manually collate it, you mentioned you have word documents so presumably you end up with like a repository of word documents and it's organised by the themes of your research that you are interested in, am I right?**

F: Yes and it's all colour coded, go through and highlight things in different colours, different chapters that I might like to write.

**I: Ok, that's very interesting because one of the things I'm really trying to do is establish a kind of generic workflow for social scientists working with this type of data so that's really helpful.**

F: I think it's a very Luddite approach that I have, I'm sure.

**I: So you've got your organised repository of information in these Word documents, my next question is: Do you use any tools to organise that or are you manually putting things in folders and labelling the folders and then you've got your colour coding within the Word documents?**

F: Yes, it's a very grounded theory approach going through things again and again so various things are coded or various themes come out.

**I: Ok, that's very helpful. So my next question is: When you are analysing the social media data, I presume that you have to just put the time in and sit there and read through it, be aware of potential context and a link to follow up things?**

F: Yeah, obviously I can't collect everything on the subject.

**I: There's a lot of richness there though, isn't there?**

F: Yeah, so what I tend to do, , I just used the first hundred and you make that clear.

**I: It's like a polarity thing there, isn't there, which would be reflected in your coding, presumably?**

F: Yeah, yeah.

**I: so if there was a way to kind of automate that type of coding, would you find that useful?**

F: Yes, I think I would. Definitely with something … again it would really have to take – people rarely make a clear statements and that's the problem. When we were looking at the material it almost needed hand coding and if you looked at that just as it was written then you think 'well, they are very pro or against but they weren't actually.

**I: That's very interesting do you use any content analysis like NVivo or Atlas?**

F: I should and I don't, I am honestly an historian and my colleagues use NVivo and they go 'oh you could do that with NVivo' and I prefer to take a small amount of the data, restricted in terms of time or something and I don't ever make claims that this demonstrates what all others are thinking about, I'm saying this is a snapshot in time.

**I: It's almost like you want to maintain the richness of the quality of your data and if you, from what you've said, it seems that you are kind of open towards some kind of automated analysis techniques.**

F: We did use something with the blogging thing which was really very useful, we found it online and it looked at the links between all the blogs and sources. That was good because you could see which were the important blogs that everybody was linking to and who was the outlier, But that was a really interesting tool and we made a lot of use of that .

**I: Was that a visualisation tool like a social network graph or sociogram?**

F: Yeah, something like that. I can send it to you but I don't know what it was.

**I: That would be great if you could, that would be wonderful. I think we've covered everything, I just wanted to ask you just a few other questions: What would you say are the limitations to your existing approach to capturing, storing and analysing social media data? I'll qualify this question: there's something which I'm discovering, a recurring theme which is our objective is not to replace social scientists with a script although some social scientists are very sceptical about automated analytical techniques from computing science because this type of data, it's generated by people, we don't have machines capable of understanding this at the moment or likely to in the near future so as you mentioned before, you do need to have some kind of manual analysis. What we're really interested in is how we can support you in the manner which, say, saves you time so you can then, for example, you mentioned your historical documents so if we could speed up your content analysis, your coding, then with that time what would you do? Would you gather more data and increase the size of your sample and then use these automated tools more or would you do more research?**

F: I think I would do more research, I'd be doing more reading of secondary sources, I'd be going off to the archives to look at other sources content that I had the basic newspaper stuff and so could go and read (*many)* books of things.

**I: So what we have in mind is a tool where, let's say you have fifty word documents and say you – well, say a hundred word documents and you analysed ten of them then the machine or the software looks at how you've coded them and then says 'within this degree of confidence I can code the rest, the other ninety with 75% confidence that I'll be coding them the same way you do' and then at the end of that process it will say, 'ok, I've done eighty, these last ten I wasn't too sure about, you'll have to do them manually', would that be something you'd be interested in?**

F: Yes, not in a quantitative way, not in a way that would then make me say 23% of the someone said this, but what it would be useful for is, for example, at the moment I want to write a paper on the physical descriptions of people so if I could get something that could pick out in blue then I could then go and I wouldn’t have to read through my thousands of words but I could go, 'ok, here's all the stuff that's in blue, I just need to look at this'.

**I: So you'd have a sub-set of your parent code and the parent code would be 'physical descriptions of and then the hundred documents from that, our tool would ideally give you 'here's a condensed version', like a summary.**

F: Yes because I wouldn’t be looking at it in a quantitative way so I would say 50% some group were talked about…

**I: Yes, it's almost meaningless, isn't it?**

F: But I'm using it as rich qualitative data to say 'look at the way they are talking about this and that would be great rather than me having to search through these thousands of words to find these few examples where they talk about the bodies. If there was some way in which that could be brought out, that would be what I would use it for.

**I: Ok, that's kind of like what we're moving towards. So what I'll do is I'll talk now briefly about some of our ideas and just bounce a few ideas back and forth.**

 **Our initial plan was to develop what we call like a 'knowledge graph' so if you go to Google and type in something like 'how do you do something in SPSS' or whatever, then it will come up with some numbers, no1: do this, no2: …I don't know whether you've seen that before, if you've looked for things in Google?**

F: Uh-huh.

**I: So what we're thinking of is we used this kind of knowledge graph approach at the point where you capture the data so we have a button, you press it and it says 'ok you are importing data from Facebook, this is your search term, this is the public profile you are looking at' or ' this is the hashtag you are looking at.' So, for example, in your case maybe you want to look –and see how people are engaging, would that be a potential use case?**

F: Yes, possibly, possibly.

**I: And within that you want to see what posts, you want to capture the posts as well as the replies to the posts, what conversations are going on and what happens is our programme pulls in all that text data and then puts it in a Word document or Excel fire for you and then it gives you a summary of it, it say 'We captured fifty posts of these posts, this was the number of followers at the time' and things like that. 'This is the number of likes' or, in the case of Twitter, 'this is the number of re-Tweets.' Again it would give you a quantitative summary but just to give you the sort of confidence in the sample or an idea of what you could do with this sample, would that be something that could be handy for you or could you see that as something not really relevant.**

F: It's not anything I've really – it wouldn’t be a priority for me, I'm more interested in the actual topics of discussion.

**I: Ok, forget the knowledge graph then. If at the point where we've pulled the data for you we then say – we provide you with the facility to annotate the data manually and then the system says to you ' I have this level of confidence that I can repeat your coding scheme for the rest of the data-set and then we could even include a library which says, if it was political, you want to detect sarcasm and irony so if we had a library that said 'we will leave it for you to decide certain cases if something sounds particularly ironic or sarcastic' and then you would manually annotate the data for a certain percent, for a certain portion of the data-set and then the system will say 'ok, I'm 50% confident' and then you do a bit more and then it says 'now I'm 90% confident', would that fit in with your work-flow an analysis?**

F: That would be something I could see us doing; pulling enormous amounts of discussions about and training it up so that it could see if someone is talking positively about the apps or negatively about the apps and then, yes, being able to say 'ok, now show me all the posts where someone is very negative'.

**I: So what types of negative arguments are made..**

F: Yes and now we can say 'they are too expensive' or we can have another code that says 'they don’t work 'and then working with that data. Yes, I could see that working.

**I: And in terms of visualisation you mentioned a tool which helped you visualise a social network graph almost, in terms of organising the content, at the moment you obviously have like a tree structure, like you would have a directory and then within that directory maybe sub-directories and named on coded themes and then the documents in there and within those documents coded annotations. Would kind of like a sort of global map of how you've organised that data, would that be something that you would be interested in or would it be…?**

F: It would be something I would be interested in, mainly because what you've just said sounds far more organised than what I am!

**I: So in that case it's potentially deterministic, it's not something that you would need, it's kind of…**

F: It's more something I would go 'maybe I could or should organise my data in that way' rather than the incredibly haphazard 'it's in my brain, I know where it is' way that I've got it organised at the moment.

**I: I think we're almost done. I just wanted to ask you just a few other questions: In terms of social media data, do you use any theoretical frameworks or methodological analysis frameworks or techniques?**

F: Not really, no.

**I: So do you have like – it sounds to me like you have a pre-defined criteria guided by the objectives of your research and that's how you are doing almost like your manual sentiment analysis at the moment?**

F: Yes, I think that's probably accurate, yes.

**I: So I'd like to show you something. This is called the 'Honeycombed Framework', this is something which I don't know too much about, I just came across it when I was going through the literature and I was wondering if you could let me know if there's anything in that which would help you organise your capture or guide you in terms of establishing objectives or potential research questions?**

F: Possibly. I suppose when I'm looking at something, I'm not looking at it as a whole and then thinking about the conversations that are going on. I'm using it more as a source of data from which I can take small groups of data. looking at the way in which people became leaders to come along and advise you

**I: The main influences (…) of leading the crowd.**

F: So this sort of thing, certainly to do with reputations from relationships would have been very relevant then. And, yeah, I can see that that sort of thing would be applicable

**I: So the last questions I want to ask you are really based on your workflow. Now we've talked about how you would capture the data in a very manual way and then you would organise it based on your themes and then code it. You mentioned that you don't use NVivo and you are doing this manually. For the record, I'm supposed to be a technical person and the way I see it the machines should serve us the way we intend to work. So there's nothing wrong with that, there's hundreds of different ways you can do this. When I started this project I found the name of our project, like I said, is tools to support social scientists, blah-de-blah and within the first day on this job I found two hundred and thirty tools in existence for doing this. So when I write the opening of my paper how am I going to say 'Well, among the two hundred and thirty tools I've discovered something novel.' But I think – when I mentioned your workflow you said you are a qualitative researcher by instinct and this is what I'm very interested in; establishing a generic, it almost defeats the purpose, but establishing a generic workflow for, not just social scientists, everyone working with qualitative data. So when we developed this tool, it's very intuitive to how you work so at the point of capture it's just 'here's your data-set and let's try to code it and then we can assist with automated methods for coding.' So I wondered whether you could speak about your workflow in terms of different stages.**

F: Well one thing that I've not really mentioned and that's probably a very important part of what I do, not the historical stuff but the other stuff, and coming from it from point of view, we would very much see the research as part of the group that are being researched rather than looking down like they are in a petri-dish and you are the lab scientist.

**I: Kind of like symbolic interactionism kind of perspective, the observer is also the observed and the whole *(hawthorn*) effect.**

F: So particularly most of the research I do involves going back to during the process and saying 'I did this survey, what do people think about it, do you want to comment on it?' Or talking to them through discussion threads about the research that I'm doing and taking you back all the time, which of course is another rich source of data because you've got qualitative comments coming.

**I: This is something I'm very interested in as well, this methodology, I use a very dry term for it because I'm a computing scientist so we're into this techie terminology stuff but I see this as almost a form as cross-validation but maybe cross-validation is too strong but it almost reinforces insights that you find so, for example, let's say I went away, wrote up my notes from the transcripts of this interview, came back to you and asked you some more questions to make sure I understand it, am I right in saying…**

F: It's more than that. It's as if I then look at your notes and start criticising them. I have found that you have to be quite robust, you are going on and asking a million people what they think of your research so far and a lot of them will say 'well, it's crap', but in itself, I've been the project that I did, I was attacked online for researching, I shouldn't be doing research on this …. There was accusations saying in fact I was this because these people would be interested in this.

**I: Wow!**

F: Yeah, that was fun. But I do firmly believe in that sort of approach. Obviously I can't go and dig up this and say 'am I getting it right?' But I think it's a more immersive way but I think it's a really important way and it's really good for impact as well!

**I: Absolutely but also I think – I'll give you a use case for how I've used Would the process I've just described there be applicable in certain areas of your research? Not necessarily as 'this is where I stop' but if part of this process could be automated to say 'this is giving you an idea of what is going on, what people are talking about', would that be helpful?**

F: Yes, that would be. And certainly because the work I'm doing on Mumsnet is very much a longitudinal study so finding out and looking at the changes…

**I: So how the polarity changes over a period…and again, would that be helpful say if you – going back to the suffragettes example, how – I don't know if you had some archives from, say, the late 19th Century to, say, I don't know, the 1930s, to see how the climate is changing in terms of what is being discussed?**

F: Yes, certainly the tone changes, Then obviously you get this overall change in the discussion which is quite difficult to pin down to one article ' it's more an overall turning against what they are doing and that's quite difficult to actually identify and probably if you have an enormous data-set then that's the kind of thing that would be helpful with that.

I: And just one last question: I think you've raised that the key limitation to all this **type of research which, in my opinion, can only be resolved by the researcher which is the issue of context. So in terms of your data sources, do you have like meta-data which gives you, say, the name of the source, who was publishing it and stuff like that?**

F: Yeah.

**I: Is there any data from social media, is there any such data from social media, data you work with, which gives you a similar idea with the context or is the context derived purely from your manual coding as in looking at the conversation that is going on rather than 'I know this user's characteristics are typically going to be against…'**

F: Yeah, I suppose I get the context from the fact that I read every single day.

**I: Again you are immersed in it.**

F: so I have that context. But it is in my head rather than anywhere else.

**I: So – sorry to keep saying 'one last question' but I think you've given me a very good idea in terms of immersiveness in working with social media data. So the way that – I've developed a small-scale application which I'm hoping to test if, at some stage, I could show you this working with, whether I could get some feedback from you, that would be great?**

F: That would be very helpful. At the moment we tend, for each different project I'm involved with, I tend to set up a blog so I tend to direct people to the blog to learn more about the project and then more about the research I've done in the past to engage with us, to do the survey or whatever and I find that that is a helpful way to approach it. It seems to be something that ticks a box for funders, they see it as a part of the dissemination process and let's not forget that practically everything that I'm talking about with social media I've got to be thinking about how will the funders see it and how it will impact. It's also very good for impact.

**I: You are engaging with the public.**

F: You are engaging with the public, you are disseminating your research and the great thing about having a blog is you can post every week about 'we've collected this now, we're thinking about this, here's a good piece of information about what we're doing', so it really is a very..it's ticking all those boxes.

**I: Would you ever consider doing this in a bit of a covert style so, for example, it's hard to come up with an example but when you put up a blog about your research and then people engage with – this is gold, isn't it, because the hardest thing for (correct me if I'm wrong) but as an academic in the nursery I see our hardest job is getting data. Now this is the good thing about social media, we've got the data so now I see these natural mergers, who are the people who can work with data, can we make these objectives meet the people who actually understand the data and that's really what I see our project as doing as a whole. So let's say the point where you've got people engaging with your blog, would you ever see yourself in a position where you would, not necessarily have to mislead people but not clearly state your objectives? For example, let's say you were trying to detect racism in certain groups or something? So I mean, the ethical approval alone is going to be quite interesting but could you ever imagine using that kind of research?**

F: Well what I tend to do is either make it very clear I'm doing research on a subject or, social media is great in that you can collect data without…you can just collect the data so discussion, I just collected the data, I didn't make any attempt to tell anybody I was doing it.. What was actually informing the whole thing is research other people had done on where they were reporting that difficultly there's well reported problems for engaging in these,

END