**I: The project is just trying to develop a tool to help social scientists analyse social media data and so this interview is about trying to get some general requirements and stuff. So I'm going to ask you about the work you are doing with social media and then what kind of tools you use and then try to narrow it down on three main areas which are 'how do you capture, how do you store, how do you analyse, what tools do you use and things like that.**'

F: I will be able to give you some insights about how I would want to use it.

**I: Yes it would be good to talk about what kind of tools you are aware of at the moment and what you perceive as challenges and, like you said, what you would like to do and we could talk about that.**

F: Exactly, that's a good idea.

**I: what would you say your discipline or subject area falls under?**

F: Environmental Psychology ….in the realms of sociology, maybe a little bit of human geography so I'm a little bit interdisciplinary in that respect.

**I: So would you consider the research you are doing at the moment within the field or the umbrella of Social Science?**

F: Oh yes, definitely, nowhere else.

**I: Yes, you would just manually pull the data, presumably?**

F: Yeah.

**I: And how would you store that data?**

F: Probably in NVivo although I have a lot of problems about using NVivo because of Macs, so they've brought out NVivo for Macs but they've brought out a kind of beta version of it, I tried it – I did the free trial of it and I hated it, it didn't have all the functionality that NVivo for PCs has.

**I: So it's a little unstable?**

F: So even though I’d waited ages for NVivo for Mac to come, when I tried it I didn't like it and I've never actually sort of got it.

**I: So you are emulating Windows and running NVivo in there?**

F: And then you can't very easily transfer the file across, you've got to go into the internet within the virtual box and put it into Dropbox in the internet in there because it doesn't speak to the Mac

**I: So with regards to social media, you mentioned Facebook so I presume that you are manually copying the text and then creating a dataset in NVivo?**

F: I haven't done that because the purpose it just them wanting to stay in touch with each other and communicate with each other and I haven't analysed any of what they've said in there.

**I: I presume you'd have questions such as 'How do researchers obtain informed consent when there's a significant volume of users to consult?**

F: Yes. Well mostly they don't so then you have to look at, especially if it's Twitter, paraphrasing, if you are going to use a quotation in a paper or a presentation or something you need to paraphrase it because if you provide that quotation word for word then if someone typed that into Google it would take them directly to that person's Twitter account so then you are compromising that person's anonymity which is a really important aspect of research ethics. It's interesting, I took this as a presentation to a conference in Europe, most European universities across Europe don't have an ethics committee like we do so they don't have to bother with consent forms or anything like that, or anything, so they were all 'this is all very interesting and I'll give it some thought' but not in the sense of 'this is something that I need to be doing.' Whereas in the UK we need to do this.

**I: That's great. So I've got a clearer idea of what your research questions are, so can you talk to me a little bit about the…can you describe the objectives of your research with regards to the social media project, you spoke about the ethics of capturing data, is there anything more you are interested in in terms of capture, storage and analysis of social media data?**

F: Just asking people how they do it and then from that we will, using NVivo or something, will analyse what people are saying and try and pull out some important themes and how they reflect on the ethics of it. But I don't have any expertise around people are storing it or using it,

**I: But that's your focus: how people are using social media data analysis tools and the ethical implications of the whole work flow?**

F: Yes, I'm not so much focused on how they are using the analysis tools, just how they are – I suppose the tools come into it in the sense of how you are obtaining the data but that's as far as we probably would go with that, it's how are you storing it, what kind of meanings are you attributing when you are analysing it to the data.

**I: So how are they coding it, essentially?**

F: Yes and are they using direct quotations in their publications and things like that and what might be the implications for people's privacy and security and things like that. But also what are the ethics of using it even if there aren’t any implications because that person doesn't know you are using it.

**I: Ok, that's very interesting. I think we've…what role does social media play in your current research activities?**

 F: Yes, I think it would be more in future research activities that I'm planning to hopefully see if I can get funding and then it would be large scale netnographies, from both sides,

**I: Organised online prejudice groups?**

F: Yeah,

**I: So it's like identifying the agenda for why people are engaging in this interaction online with these different organised groups?**

F: Yes,

**I: It's interesting , it gives me an idea of the objectives.**

F: So the big-scale analysis I would do of social media data would be around all of those kinds of things.

**I: So at the moment if you assume that you, if you have some social media data to analyse for these groups, what tools would you use for capturing it?**

F: Yes, I don't know. I'm sure there are tools out there.

**I: Yes, it's just that I ask everyone the same – I try to keep it consistent with all the other interviews..**

F: That's fine. Well, for example, if I was going into Twitter I might start by doing a hashtag search and maybe just use the tools that are in there, I would have to Google or ask colleagues how do you do a proper netnography.

**I: So in terms of establishing a workflow for your activities you then search for hashtags through Twitter and then when that returns some results would you then, for example, as an initial starting-point would you try to save the web page maybe or something along those lines?**

F: Yes, that's a good idea so you could save the web page or maybe copy and paste the Tweets that were relevant to you but I can see that gets problematic when you are looking at larger numbers of Tweets.

**I: And then I presume, again, keeping this at the initial way of doing things you would then import that into NVivo and just…**

F: Yes but you would have to do it manually, wouldn’t you, by copy and paste, copy and paste and that would be a nightmare. So something that could search Twitter or Facebook groups because I have no idea how you search Facebook groups so a tool that could do that would be amazing. But I don't know whether you need to work with Facebook to develop that kind of tool, I don't know, I don't understand that side of it. But then something that can import the results into NVivo or put it into some kind of format which you could then easily…NVivo needs to understand that – so if it just imports the whole lot as one body of text, my understanding of NVivo, if you put that into one file that's almost like understanding it as one person. So how do you then divide them all up in a way that NVivo understands, (as) separate people.

**I: There's a tool for NVivo specifically, called NCapture, and what you can do with that is you can go to any public profile or public Twitter account and you can pull as many Tweets or posts as you like, and what that does is it brings the data in as a dataset so it's essentially like a spreadsheet so each row will represent one post or Tweet and the columns will represent various types of information. So, for example, who the person is, the content of their Tweet, does it contain a link, what is that link. If someone has geo-tagging enabled it will have that as well. As an initial analysis it will give you a cluster based on how many people have re-Tweeted it based on similar word frequency and things like that also the geo-tagging, it will show you a map showing where people (*have pointed at)* and you can interact with that map and click on certain clusters within the map, it's just dropping pins.**

F: That sounds good, I'll have to check that out, yeah.

**I: Yeah, I can show you that if you want, later.**

F: But my biggest point would be that a lot of the – communities are happy to use Facebook, less happy to use Twitter so is it possible to develop a similar tool for Facebook groups, public Facebook groups?

**I: Yes, the same tool would work with public Facebook groups. If the group or like a profile is not public then you can pull a PDF off the page and you can do that with any webpage and similarly with a blog post although I haven't tried it with that but I can show you because it's quite satisfying when you see the tool just pulling all the Tweets and it counts them as it goes and then you have it in a nice data-set, it's already classified in terms of…**

F: Can you also use that same tool to search by a certain word, key word?

**I: Once the data is in NVivo you can search how you would normally.**

F: But can you search it in Twitter in the first instance?

**I: Yes so what you could do is, as we were talking about, you could look for a particular hashtag and then you could pull all the Tweets related for that hashtag and then that would be in Nvivo and you could start doing your manual coding. Similarly with Facebook, you could say 'right, this is a person or a group, a public group or person I'm interested in and I want to pull all their posts' and it does give you the option to either include or exclude people with Twitter, whether they are re-Tweeting or not. That's good. So there's one thing I wanted to ask you which was about whether you use any theoretical or methodological frameworks when you are considering social media. So do you use any kind of theory or methodology which defines features and characteristics of social media which you are interested in?**

F: Go and sort of like explain that to me again**.**

**I: For example, would you approach a public Facebook group or select a public profile group or Twitter account and then try to identify whether the users provide information which reveals quite a lot about themselves? Or for example, the extent to which users relate to each other. So for example, if you have a Facebook profile with someone who has quite a lot of connections, would you then look at the frequency in which they are communicating with each other to see what the strength of the network is? You mentioned netnography?**

F: Netnograpny, so is what you are talking about maybe social network analysis or something like that?

**I: Yes but before you do the analysis do you have any kind of criteria…**

F: To try and scope it out?

**I: Yeah.**

F: I have to admit, I haven't, because I haven't done those analysis yet.

**I: So any kind of information that you have about the behaviour of the groups and establishing a criteria for further analysis would be derived from those interviews?**

F: So the findings from the in-depth qualitative interviews would inform what were the interesting areas to pursue. But yes, I would be interested in trying to…it's a difficult thing, you can't really understand how to pin it down like how many connections do they have and how they are using those connections, that's not something I can find out because you can't tell from looking at a person's Facebook page There's difficult things to really get to and you can only really get to that through interviews, I think.

**I: So there's a limit on how you can establish identity and status as are they the subject, focus of your research, you can only really do that through interviews rather than…?**

F: Yes, because there's ethical implications again from nosying around people's Facebook profiles, for example, isn't there? Which with Twitter there is not so much because people kind of know their profiles are public, don't they? So you can look at and see how many friends a person – followers and how many people they follow, on Twitter it's all open unless maybe there's a way of closing it. But on Facebook some people have that information readily available, some people have it locked down so you can't even look to see what they are friends with, everybody has got very different levels. So from that point of view I think it makes it quite hard.

**I: So I'll just review some of the things we talked about. So you mentioned NVivo would be your primary tool for storing, well for potentially capturing but definitely storing and analysing your data. When it comes to analysing.**

F: To be honest I might even sort of – maybe it's just really old-fashioned it's just that I can easily get to things without having to open up NVivo and I might have, especially with the trouble I've had with NVivo I might even copy and paste stuff, just into Word files, it's very old-fashioned I know but I might do that actually.

**I: Well I think if you have interviews transcribed you then import them into NVivo…**

F: So that's the way you would do that, yeah, and that's the way I've always worked so there's an element of…yeah.

**I: So the types of social media we've looked at are predominantly Facebook and Twitter but you mentioned blog posts as well?**

F: Yes and I suppose blog posts are sort of public as well, aren't they? So there's no reason why you couldn’t look at blog posts and also look at replies to them which….I do have quite a limited understanding of these areas but that's what Ruth did, wasn't it?

**I: And one other thing I wanted to as you as well, is you mentioned some of the limitations of social media data and how the actual data you are interested in, you need to find this out through in-depth qualitative interviewing…**

F: For this stage of the project, yeah, so if I was to get further funding it would be different.

**I: Could you see this as a way of triangulating any findings you get from social media data, aligning it with your interview data?**

F: Yeah. it's difficult, isn't it, in order to be able to identify people as travellers they either need to have said that they are, you can't just assume that they are even though there are some characteristics which make them stand out or be in one of the groups, like I say – some of them are closed and there are groups from the other side of it, like the attitudes towards communities, yes, then I would be able to access that material much more easily I think.

**I: At the stage where you are going through your interviews and you have them in NVivo, do you use the nodes to code your data and stuff?**

F: Uh-huh.

**I: So what we're interested in developing is a tool which initially gives you a summary of the data you've captured to support you with your analysis, so I'll give you a use-case. When you pull a load of Tweets from Twitter or posts from Facebook for a group you are interested in, we'll provide you with a summary saying this is the extent to which these Tweets or these posts are liked or re-Tweeted or favourited and maybe some summary statistics like talking about the demographics from where people are based, just descriptive statistics so percentages of where people have posted or re-Tweeted and differentiated based on what information we have from the dataset to give you a summary of the data prior to you starting your coding, like of the data-set, would that be beneficial or would it – do you think it would introduce a bias to your analysis?**

F: To the analysis. I think it could be beneficial as long as it's something that could be turned on and off so that the researcher can decide whether or not that that is valuable to the research question. So if it's about the extent to which particular attitudes are becoming normalised or embedded in certain group opinion then that would be really helpful because being able to see how many people like or share something gives you a good indication of…

**I: But it wouldn’t be coded, it would just be like this….it would be globally looking at the whole dataset so it would be like 'there were this many Tweets focused in this part of the country, this number of them are re-Tweeted and here's the distribution here geographically for here for what…'**

F: But it's the same, even if it's not about one Tweet, even if it's about a group of Tweets, if you are looking at a particular theme then that essentially is quite important.

**I: Like a hashtag, say.**

F: Yeah, especially if it's something that's localised but has got a geographical spread then it can tell you something quite interesting. So I think as long as it can be turned on and off that would be really good.

**I: There's one other thing I wanted to ask you about which is the process of coding your data so let's assume it's either interviews or social media data. If you had a tool which, when you've coded a certain portion of the dataset and you are saving your model and your codes, if the system could then say to you 'within this degree of accuracy I can proceed to code the rest of your data', would that be useful?**

F: Yes, really useful. As long as maybe it could do ten or twenty or something and let you look to see how it's doing it so you actually believe in it and trust it. You would need to start the process off so it could see how you were doing it and then if it can use machine learning to carry on with that, as long as you are able to check it, you wouldn't want to check all of it because you might as well have done it yourself but if you could just check, say the first twenty or thirty.

**I: Some form of cross-validation then?**

F: Yes, exactly.

**I: And how about the facility to be able to modify the way that you are coding the data because you may find something in the second portion of the data which then you are aware will exist in the first part so you need to go back.**

F: Yes if you can do that, that would be brilliant. You don't have the time always to do that. You just realise that 'this is a really interesting area that's emerging, I'll have to go back.' As long as, again, it learns from how you are coding that particular node or whatever you want to call it, yeah.

**I: And the last question would be: What would be your dream research tool for using social media data?**

F: Hmm. Well essentially just the best bits of everything that we've said so NCapture sounds really good and I haven't used it yet so something like that; you can use it based on search terms and it brings up these thematic clusters of Tweets or Facebook posts that it can then import them into NVivo for you in a format that makes it easy for you to kind of code them and then you start the coding procedure and it learns from what you are doing, you check that you are happy with what you are doing and it does it all for you. That would be the dream. That would be the dream because then you are not going to be spending the traditional way of using NVivo – to spend months…and there is a way of automatically coding in NVivo at the moment but it's not very good. And there is actually something about the researchers still doing some of it because you get such a sense of your own data and an understanding of it that you wouldn’t get if the whole thing was automated so it's still important to have some kind of hands-on part of the process. I think it changes, something like Grounded Theory Approach, for example, you couldn't have that fully automated because part of the actual analysis approach is the researcher going through it multiple times and embedding themselves in that data and immersing themselves in the data, as they say. So I think you can't take that away completely because I think analysis would become a little bit superficial.

**I: There needs to be a clear boundary between what the machine is doing to complement and support the researcher…**

F: But not completely replacing that human element to it, yeah.

**I: Because the data should be interpreted by the person rather than the machine.**

F: But then of course the human is going to come in, the researcher is going to come in and look at the different clusters of codes or nodes or whatever, so that – you would hope the person comes back to that and then takes the analysis from there. There's only so much that the machine can do or the system**.**

**I: And would you be happy with the boundary between what the machine can or is allowed to do based on this descriptive summary of the data-set?**

F: Right, say that again.

**I: As you were saying there needs to be some kind of manual control retained by the researcher because it should be interpreted by the researcher. But the boundary between what the machine is doing automatically for you, would be comfortable if that was just providing you a summary with some descriptive statistics and describing the dataset for you. You don't think that would compromise you or bias your analysis?**

F: I see what you mean. It might…yeah, I see what you mean now, it might actually cause you to understand the data in a particular way and once that's in your head…tricky..

**I: If it was just showing you trends in the data, so re-Tweets, you've already selected the…**

F: Ok, so it's not giving you a theoretical model and saying like 'issues of trust are more important than practicalities of x, y or z' ?

**I: No.**

F: No, it doesn't bias you then, in my opinion. If it was starting to try and give you some kind of theoretical model, that's maybe in the future, that's maybe some kind of science fiction but once it starts to do that then yeah, you are going to be bias because you think 'I already have the model, I don't need to come up with that myself.'

**I: But just to clarify: if it gives you this summary description of trends in the data but there is no coding, there's no content done yet because that is still retained by the researcher, that would be ok, in addition to the facility to turn that on and off?**

F: In theory I think that would be ok, yeah. Seeing it in practice would make it easier to understand….I think it sounds ok, yeah.

**I: I keep saying this is the last thing I'm going to ask you but it gets very interesting the more you go on. With regards to this knowledge graph, let's say you import the data, you turn it on and you see – here are some trends, some re-Tweets, re-posts, there's quite a lot of images, video and so on so when you start going through the data and you are manually coding it, once you've completely annotated the whole dataset, would it be – again, it's a question of whether this would compromise your research, if you then re-looked at the knowledge graph and it incorporated into its trends and summary report, your codes and then gave you another….?**

F: No, I think that's fine. And then gives you – I think it's helpful because those are your codes so it can't bias you just by incorporating those into what you saw previously. I think that's ok, yeah. Because what you want to do, you are a qualitative researcher so you are not supposed to be looking at the numbers of and the frequencies of but you are still looking for patterns so you are still looking for this theme has emerged as a stronger theme, for example, and although they made me do it in my thesis, I try not to report on numbers of, mentions of this and that, personally – it's just all personal taste. But I try not to report that but it's still interesting for me to see that because it gives me a feel for what are the more important areas and that might have been biased anyway if you were doing an interview it would have been biased because of what questions you've asked. But social media analysis, it's going to be biased by the search times that you use or whatever but having said that, it's what comes out within that, isn't it? So I think applying that to some kind of knowledge graph would be really good but I'd have to see it in action to know, mock-up something.

**I: Well thank you very much.**

**END**