Netnography & Demography: Mining Internet Discussion Forums on Migration and Citizenship

Pablo Mateos, CIESAS / University College London
Jorge Durand, Universidad de Guadalara / CIDE

Abstract

This paper explores the potential value of netnographic methods in social media applied to migration studies. It develops a pilot netnographic analysis of an internet discussion forum on migration and citizenship, in particular access to Spanish citizenship. 54,920 messages posted by 6,813 persons were automatically downloaded and classified into a database. Manual classification methods were performed assigning user profiles to participants and establishing key themes and 'citizenship trajectories'. Through quantitative and qualitative analysis we identify a series of strategies to access Spanish citizenship that seek to maximize migrant's possibilities given their life stories. An unequal pattern in the 'geography of access' to Spanish citizenship arises from these results. Many of these findings were previously absent from the migration and citizenship literature. The next research frontier in netnographic methods is to automate the task of assigning user profiles in discussion forums and the identification of key topics. To encourage new research in this direction, the paper ends proposing new avenues for research by adapting text-mining techniques from computer science and natural language processing to netnographic methods. Over the coming decade, promising new research developments in this area will revolutionize traditional population research methods, most definitely in migration research.

On-line research methods in social science: netnography

Academic research of people's views published on the social web, comprised of internet forums, blogs, *twitter*, *Facebook*, *YouTube*, comments to news sites, and the like, has rocketed over the last five years. According to various authors these social web technologies can open up a new era of social science research (Golder & Macy, 2012). '*Analogous to what it must have been like when they first handed out microscopes to microbiologists, social scientists are getting to the point in many areas at which enough information exists to understand and address major previously intractable problems that affect human society*' (King, 2011, p. 719). One of such problems is the study of human migration, a complex rapidly changing phenomena that straddles across the social container of the Nation State, and hence escapes most conventional forms of social research methods. There is a need for new empirical research approaches that focus on migrants' agency in order to understand transnational migrants' complex practices.

The impact of contemporary communication technology on international migration, especially in facilitating transnational activities and diasporic family and social relationships across the globe, is a fast growing topic of research (Miller, 2011). The Internet is also quickly becoming 'a site to conduct fieldwork' in social sciences, because of its growing importance in reflecting personal relationships and life experiences (Hesse-Biber & Leavy, 2008). Hence, carrying out research online has become an efficient way to collect qualitative evidence of rapidly changing social and behavioural processes (Kozinets, 2010). However, research into online research methods in

migration studies is very scarce, not to mention citizenship studies, the focus of the latter part of this paper.

One of these new on-line research techniques in social science is *netnography*, defined as 'participant-observational research based in online fieldwork'(Kozinets, 2010, p. 60). It is a broad term that encompasses the observation of people's views published through online communities, on internet forums, blogs and social media and analysed through an ethnographic perspective. The main advantages of netnography, are its low cost, and the ease of access to a large volume of participants' material compared with traditional survey methods. '*Netnographies […] can also hone in, narrow, and focus on particular relationships or previously identified constructs'*(Kozinets, 2010, p. 80). The obvious disadvantages are lack of control over the sampling framework, as well as the absence of a questionnaire, interview schedule, observation plan or any other research structure and strategy that could be devised prior to data collection. These disadvantages may introduce bias into the results, especially critical when attempting to generalise from the results. However, we judge that the considerable advantages of this method clearly exceed its drawbacks, allowing researchers to collect a very large amount of opinions form a diverse range of participants without any 'researcher observation effect' and at a fraction of the cost of alternative methods. Therefore, if netnography is seen as a complementary and exploratory methodology, then its considerable advantages exceed its drawbacks.

However, netnography is still in its infancy and relies on manual extraction, reading and classification of user messages. It is rather surprising that the first textbook on netnography (Kozinets, 2010) actually recommends browsing individual messages from websites and manually 'cutting and pasting' those of interest of into a word processor. This is partly because most new developments in mining the social web have come from the computer and physical sciences, and also because it is much harder to extract meaning from unstructured qualitative opinion data than extraction of quantitative measures and binary sentiment analysis. Most breakthroughs have come by collaborations between social and computer scientists (McCormick, Lee, Cesare, & Shojaie, 2013).

This paper is situated at the frontier of this research challenge. It aims to developing new ways to extract qualitative social opinions from the social web, in particular applied to the study of migration and citizenship.

Internet discussion forums on migration and citizenship

Internet discussion forums on migration promote the free exchange of information about the cumbersome process required to migrate to a country, obtain legal status or access citizenship through various routes. These forums gather an extremely wide range of evidence about migrants' mobility and citizenship trajectories and inter-generational life histories, documenting every step of their struggles, uncertainties, general concerns and pointing out policy contradictions from person or household level experience. For example, the following quote from one of these forums is very revealing of Castles' (2005) hierarchies of citizenship, and the pragmatic value of a European Union (EU) passport for global mobility: *'One of the reasons for me trying to get my Portuguese passport so desperately is that I'm an IT consultant who travels the world trying to market my skills, and as of recent I'm finding that traveling with a South African passport is prohibiting me from obtaining work.'* (John, South African with Portuguese grandparents, 2008). This message elicited 305 responses from other users, so the experiences or opinions of other people in relation to this topic can be traced thematically (*message thread*), as well as following the stories of each of those individuals through other messages in other discussions (*user history*). But how can these messages be detected, retrieved, parsed, and classified? Even more problematic, how can meaning extracted

its unstructured, informal, context-dependent and always incomplete text? This paper attempts to make a small contribution in this direction, pointing towards potential avenues for future research in this promising area.

This is the right moment to conduct netnographic studies of social web content on migration. First, vast amounts of self-reported material on personal experiences on migration and citizenship have now accumulated on the web over the last decade. Second, semantic web and text mining techniques that have been developed outside social science are now mature enough to be implemented in migration and citizenship research. Third, there has been a qualitative shift in the way the internet is used by transnational communities, facilitating movement as well as long-distance mundane interactions. Fourth, growing interest in genealogy facilitated online, combined with the increasing value of EU citizenship and the *ius sanginis* paths maintained by some countries, have made the web a unique forum for sharing genealogical information. This may build a person's ethnic capital, with the support of other forum members, to a point where it is capable of being transformed into EU citizenship. Finally, we here focus on the study of Latin American migrants, a collective we selected because of: a) the practice of multiple and external citizenships is common among mobile Latin American migrants; b) most Latin American countries present high internet penetration rates in the context of developing countries; c) although researching virtual communities can be obfuscated by language barriers, most Latin American migrants speak (and write) Spanish, an official language in 22 countries, easing research in a broad range of countries with a single set of natural language processing tools.

The aim of this paper is to develop a prototype of netnographic analysis of a migrants' internet discussion forum on Spanish citizenship, with a view of setting a research agenda for future semi-automatic netnographic methods on other discussion forums. In particular, here we focus in investigating how migrants' legal and mobility trajectories are made up of sequential migration and citizenship decisions, of a rather pragmatic nature following adaptive behaviour. This view of citizenship, based on migrants' agency and empirical 'bottom-up' research, is not common in the citizenship and migration literature, which tends to focus on normative issues and official national categorisations and statistics. We do so through a combination of quantitative and qualitative analysis of an internet forum on access to Spanish citizenship.

Methodology

Web harvesting

The internet forum analyzed here was called 'Registro Civil' (Civil Registry) (http://groups.msn.com/registrocivil), it was created in May 2001 and was active until February 2009 when Microsoft deactivated the discussion groups platform on which it was run. The original objective of this internet forum was to facilitate the exchange of information for foreigners interested in services related to the Spanish Central Civil Registry, which deals with civil registration records of Spanish nationals living abroad as well as naturalization registrations. However, in a short time it became the main forum to obtain key advice on legal aspects related to immigration and nationality procedures in Spain. The vast majority of messages with queries and their replies were posted by the forum users themselves (i.e. not experts in migration), developing a very efficient migrant self-help system. These users shared their experiences and key information related to all sorts of migration administrative procedures in Spain, mainly the process of migrating to Spain, requesting or renewing residence permits and acquiring Spanish nationality. The

messages' content was publically available on the open internet, what made them visible to Google searches on migration or nationality issues, what in turn contributed to increase participation and dissemination of its contents. In order to send messages users were required to obtain a Microsoft passport account, although this was not required to read other users' messages. This forum comprises an ideal resource for research purposes because of various factors; i) the sustained length of time during which the forum was active (2001-2009), ii) the fact that this activity coincided with the highest period of immigration in Spanish history, and iii) that the confidentiality of participants is further ensured since the forum content had been removed in early 2009 after the closure of this service by Microsoft.

A simple computer program was developed in Java to repeatedly access the internet forum and copy each message's content. The program followed a specified sequence; 1) it opened the first message ever sent to the forum identified by the number at the end of the URL address in the message board, 2) it read the message's content in HTML format, 3) it copied its content as an HTML file, 4) it extracted the relevant fields (message ID, sender ID, date, title, text, etc), 5) it stored these fields into an Oracle database, and 6) it repeated these steps 1-5 iteratively for all messages in the forum until reaching the last message. In this way the entire content of the forum spanning across seven years, from May 2001 to May 2008 (the time when the contents were downloaded), were stored in a local Oracle database, comprising a total of 54,920 messages. These messages were organized in a database according to some basic characteristics.

Database structure

The entire database comprised a total of 54,920 messages posted by 6,813 persons (or more precisely, unique user ID names). Each message was posted to one of five possible discussion boards according to the message's main topic (figures refer to percentage of messages per discussion board): General discussions (36.9%), immigration permits (30.6%), nationality (30.7%), current affairs (1.8%), and civil servants (0.02%). The analysis presented here focuses solely on nationality issues and therefore only the characteristics of people who posted one or more message to the 'Nationality' message board were analyzed. This sub-collective was comprised of 2,860 persons, representing 42% of the total number of the forum's users, and, unless otherwise specified, constitute the base population to which all figures in the paper refer to.

The first stage in the analysis consisted in capturing a set of basic reference indicators per user in order to classify them according to general characteristics. Having a classified database of 2,860 users makes it much easier to then move on to analyze the content of their tremendously large volume of messages (41,904). The results presented in this paper primarily focus on the quantitative evidence collected by classifying user profiles and describing the main types of experiences reported in the forum, while qualitative quotes are used to illustrate the main points.

A total of 16 variables for each person were manually captured by a research assistant (see their description in Table 1) related to the following aspects; a) country of origin, b) presence of Spanish ancestors, and c) migration status and experience (details about the migratory process, residence in Spain and access to nationality). Some users offer all sorts of details about their life histories, migratory experiences and the process of accessing nationality, and that of their relatives, while the great majority only report the minimum pieces of information required to interpret their question. Therefore, the classification of users using these 16 variables relates to the 'lowest common denominator' aiming to maximize coverage rather than gaining in-depth information about a few users' experiences. The data capture of these 16 variables for the 2,860 users was performed in a simple desktop database system using a bespoke menu that facilitated browsing through each user's messages and at the same time the manual data entry screen. The system allowed viewing all

messages posted by each user, which were grouped by message board (nationality coming first) and then ordered in ascending order by the message's date. The analysis was performed manually by a research assistant, who for each user read his/her message/s in Spanish until all 16 variables were completed or the message content was exhausted, repeating this process for all users. This was an intensive and arduous process that enabled the classification of 2,860 people through reading the content of all or part of the 41,904 messages they posted, with special attention paid to those 16,856 messages sent to the nationality board. Attempts were made to automatically classify user profiles by country of origin, and some of the other variables mentioned above. A dictionary of common expressions and a set of speech disambiguation rules were built, but initial progress, although successful, was too slow for the resources and time constraints available. Therefore, in the results presented in this paper we finally adopted a manual approach to message and user profile classification. Developments carried out on the automatic classification have been documented with the aim of facilitating further research in this area. These are discussed in the future research gaps and challenges section of this paper.

Analysis and results: Case study on Spanish nationality

Of the 2,860 persons analyzed, the user's country of origin was identified for 1,596 persons (55.8%), amongst which Latin American countries clearly predominate (1,416 persons or 88.7% out of those with a known country), especially Argentina (36%), Venezuela (13%), Cuba (7%) and México (6%) (see Figure 1). The variable of presence of Spanish ancestors was compiled for 1,366 persons (47.8% of the total), of which 81.2% declared Spanish ancestry while the rest made it clear that they did not have any Spanish ancestors. The rest of the users do not declare any ancestry information (52.2%) but we would expect that the great majority do not have Spanish ancestors, since this situation would clearly influence their possibilities of acquiring nationality and it would be a fact worth to mention in any request sent to the forum.

Figure 1 shows a histogram of the number of users per country of origin classified by the proportion of people with Spanish ancestors. This figure is useful in understanding the wider geography of interest in Spanish nationality, as reflected by this forum. The aforementioned top four countries (Argentina, Venezuela, Cuba and México), not only present the largest number of total users, but also the largest proportion of people with Spanish ancestors, clearly dominated by Argentina. These four countries were the largest recipients of Spanish emigrants during the first half of the 20th century, and therefore, where we would expect to find a larger number of recent descendants of Spanish nationals interested in acquiring Spanish nationality via ancestry provisions. Moreover, Figure 1 also shows a large proportion of Spanish ancestors within users originating in Uruguay, Chile, Brazil and the United States, while the proportion is much lower amongst those originating from (in descending order): Colombia, Peru, Dominican Republic, Ecuador and Morocco. This 'nationality gradient', from the historic destination countries of Spanish emigration towards the origin countries that have most recently sent migrants to Spain, is identical to that shown in official statistics (Instituto Nacional de Estadística, 2012) and thus further validates our sample obtained through the internet forum.

Drawing from this evidence we can conclude that in general terms the countries of origin heavily represented in our database are those with a larger presence of descendants of Spanish emigrants during the 20th century. This is not necessarily a surprise since we have only taken users that posted messages to the nationality board. The aforementioned exceptions (Morocco, Ecuador, etc.) simply reflect those countries with an important volume of contemporary immigrants in Spain, but not historic Spanish ancestors. Furthermore, it is worth highlighting the absence of Bolivia, a

country of origin of an important volume of immigrants over the last decade, but whose representation in the internet forum is extremely low. The Bolivian absence and the low number of users from Ecuador, reflect perhaps a much lower penetration of Internet usage of these migrants, both in the origin and destination countries, a situation that is also related to social class differences between national migrant groups in the flows to Spain. In fact, these two countries present the lowest internet penetration rates in Latin America, 2.1% in Bolivia and 11.7% in Ecuador compared to a regional average of 23.6% (The World Bank, 2011).

Nationality application and acquisition process

A great majority of forum users (2,202 or 77%) have obtained, applied for, or are in the process of applying for Spanish nationality, of which 22% report having already acquired Spanish nationality. However, we should expect that the majority of those whose application was successful will not come back and report their outcome in the forum, and thus the overall application success rate is probably very high. Moreover, 176 people or 8% of those who applied for nationality, mention the year of initial application as well as the final resolution year, so that their average process waiting time can be calculated, being 1 year and 9 months. The data has the following distribution: 3% less than a year, 25% one year, 64% two years, and 8% three years. The geographic distribution is not significant because of small numbers in this sub-sample.

Furthermore, 28 of these users also declare their year of arrival in Spain as well as the year of nationality application. The difference between the two dates thus represents the average waiting time that these users lived in Spain before applying for nationality which is 3 years and 3 months, and the mode is 3 years. In 22 of these cases the year of acquisition of nationality is also known, so the total time between arrival in Spain and acquiring nationality is 4 years and 3 months. However, these figures must be taken with care, not only because of the obvious dangers of generalizing from this small self-selected sample of our population, but also because the dates were reported and captured as single year units, not in months or days and hence all averages should be read within confidence intervals of about 6-12 months.

Bearing in mind these caveats, the typical pattern followed by migrants is that after arrival to Spain they wait between two and three years before applying for nationality, a process that takes another two years to resolve resulting in an average period from arrival to becoming a Spanish national between 4 to 6 years. On top of this time, qualitative analysis of the messages actual content shows that in many cases after the nationality is granted many users have to wait between one and two years before their new nationality can be registered in the civil registry and a passport can be issued. In many cases this long wait creates a legal limbo with respect to their migratory status, as we discuss in the 'irregularity' section.

Residence in Spain

Information about residence history in Spain is available for 61% of the 2,202 users that have applied or obtained Spanish nationality, while the rest (39%) declare having always lived in their country of origin or elsewhere outside Spain. Moreover, combined information on country of origin and residence history is available for 1,125 of these users, and hence we can calculate the percentage of persons per country that live or have lived in Spain. In descending order such 'rate of residence' per country is: Morocco (90%), Colombia (85%), Ecuador (79%), Peru (58%), Dominican Republic (58%), Cuba (55%), Venezuela (51%), Argentina (45%), Mexico (45%), Brazil (41%), Uruguay (38%), Chile (38%), and United States (31%). This list of countries is somehow the inverse of that shown in Figure 1, where some of the countries that have the largest number of messages and

Spanish ancestors (Argentina, Venezuela, Cuba, Mexico) appear here with lower 'rates of residence' than countries such as Morocco, Colombia, Ecuador, Peru, or Dominican Republic. Morocco, is a symptomatic example of this relationship, it appears in the first position on the above list, with a rate of residence in Spain of 90% while in Figure 1 appears in the last position with only 18% of persons with Spanish ancestors.

This relationship between ancestors and residence is clearly established in Figure 2, a scatter plot comparing for each country of origin the rate of residence in Spain with the percentage of users with Spanish ancestors (rate of ancestry). Two distinctive groups can be identified; one on the top left section, with a high rate of residence in Spain and low percentage of Spanish ancestors, and another in the bottom right section, with the reverse characteristics. The arrangement of countries in this graph clearly exposes the relationship between these two variables. Such relationship is probably the result of preferential treatment in the nationality legislation towards people with Spanish ancestors who are not required to reside in Spain in order to acquire Spanish nationality. These users post messages in the forum from their countries of origin, with the intention of applying for nationality through the ancestry route, while the rest of users write after having become residents in Spain. This omnipresent dichotomy; ancestry vs. residence, explains the asymmetric geography of access to Spanish nationality discussed in the previous section, which will become even clearer when analyzing the type of application for nationality in the next section.

As regards to declared intention to return to the country of origin, this information is only available for 96 users. 58% of them have the intention to stay in Spain, 19% have already returned to the country of origin, 8% intend to return soon, and 15% come and go in circular migration movements. No significant differences are observed by length of residence or year of arrival. However, it is surprising to find several users who fear losing their newly acquired Spanish nationality if they decide to return permanently to their country of origin, and others who assume automatic links between nationality, residence and access to welfare benefits:

Hi, I am Colombian and I have Spanish nationality with DNI [(national identity card)] and everything. I now live in Colombia and my question is, can I lose my [Spanish] nationality for living in Colombia and not in Spain?. If so, what do I need to do not to lose it?. If I return to Spain will I have problems for not having had National Insurance contributions, or as a Spaniard I won't have any problems? (Claudio; Colombian, return migrant living in Colombia; 2007)

Typology of routes to nationality

The legal route to apply for nationality is known for 73% of the 2,202 persons who applied for it. Eight different possible routes to nationality are captured in the database, amongst which four of them clearly predominate amongst users (93.3%); *a*) ancestors – grandparent/s (15.4%); *b*) ancestors – parents (20.9%); *c*) spouse and children (23.1%); and *d*) residence in Spain (33.9%). If the two ancestors categories are grouped together (*a* and *b*) we can then distinguish three broad types of routes to nationality; ancestors, spouse/children and residence, which form the basic typology of access to nationality discussed here.

The analysis of the different routes to nationality broken down by country of origin further reveals some more interesting patterns, especially between the two types of family related routes; ancestors and spouse/children. Table 2 shows a list of 19 countries with five or more people for which the route to nationality is known. Each row reports for each country, the percentage of users that applied for nationality through the ancestors or spouse/children routes, as well as the difference between the two percentages expressed as a quotient from -1 to +1 according to equation [1]:

$$\text{Difference quotient} = \frac{s - a}{s + a} \qquad [1]$$

where *s* is the percentage of users per country that chose the spouse and children route, while *a* is that of the ancestors route. This quotient permits to classify the 19 countries in three clearly differentiated segments; those countries where the most predominant method is the spouse and children route (top segment), those where the ancestry route predominates (lower segment), and those where the mix between routes is more balanced (middle segment). Once more, these three groups of countries follow the same aforementioned patterns when comparing rates of residence in Spain with percentage of Spanish ancestors, as summarized in Figure 1 and Figure 2. However, a new dimension can be appreciated here, since in the segment where the spouse route predominates we can observe Eastern Europe countries, Morocco and four Latin American countries without recent history of Spanish emigration; Dominican Republic, Peru, Colombia and Ecuador. These are all countries with higher rates of intermarriage with Spanish nationals, a pattern that in the four Latin American countries could be determined by the feminization of the migration flows to Spain, given that in general terms women have a higher propensity to marry men from the destination country than vice-versa (Durand, 1998). In the segment dominated by the ancestors route, we find the countries with a history of Spanish emigration mentioned before, plus Western Sahara, that although with small number of cases reflects the importance of the recent Spanish colonial past in this type of route to nationality.

Irregularity

The analysis of irregular migration status amongst users is rather restricted, since only 58 persons declare to have been in an irregular situation at some point. On first sight it is rather surprising to notice that 49 of them have actually applied for Spanish nationality or intend to do so, and even four of them report having already obtained the nationality. However, upon close inspection of their messages it is clear that these are persons that have switched between migratory statuses, either from irregularity to nationality via marriage or one of the regularization programs (amnesties), or from regularity to irregularity while they wait to obtain the new Spanish passport and their residence permit has expired. In most cases the legal situation of the families concerned gets very complicated, because some family members have already acquired Spanish nationality while others are in the process and sometimes become irregular because of the lengthy waiting times. This is the case of Marisa that clearly shows the fine line between being an irregular migrant and a national, something not always discussed in the citizenship literature:

My case is as follows, I got Spanish nationality (great!) after this I requested the nationality for my two daughters but after a year and four months I have no news at all. […] One of them is now 18 and her residence permit has expired for several months […] I am desperate and feel powerless because my phone calls and enquiries are going nowhere and I cannot find a solution. My daughters have been living in Spain for 7 years and one of them needs to work but can't because she has no work permit. The nationality will be the solution but we don't know anything about how the process is going or how much longer we have to wait. If anyone can answer my questions please help. (Marisa, Ecuatorian; 2007)

In some of these cases a complete history of family regularization and naturalization can be traced longitudinally by following messages posted to the forum over time. For example, one Cuban woman writes in June 2007:

I am Cuban and I am as irregular in Spain. My mother is in Cuba and has just obtained the Spanish nationality because her father was Spanish born in Spain, that is, she was Spaniard 'by birth'. I have requested the residence permit for being the daughter of a Spaniard by they have rejected it, what can I do? (Gladys; Cuban, resident in Spain; 2007)

A year later, after having acquired Spanish nationality she is interested in passing it on to her

husband:

Hello: I am Cuban and I have Spanish nationality. My husband is in an irregular situation. I would like to know if he has the right to regularize his situation in Spain if we have been married (in Cuba) for 8 years. Regards (Gladys; Cuban, resident in Spain; 2008)

This is an illustrative case of how access to nationality through ancestors allows to regularize the migration status of persons from several generations (the daughter of a Spanish emigrant, a granddaughter and her husband) living in several countries (Cuba and Spain) and regardless of irregularity in migration status.

Marriage as a migration and naturalization strategy

Some messages also reveal how marriage with a Spanish spouse is used as a strategy to migrate to or remain in Spain (regularization), while some persons carefully weigh the cost that a divorce could have in the process to access Spanish nationality, as reflected in the following quotes:

I am Ecuadorian and I have been living in Spain irregularly for over 4 years. My boyfriend is Spanish and we are getting married on the 17th. […] I would like to know if I can go on a trip abroad for my honeymoon the day after the wedding […]. What documents do I need to exit and enter Spain? Thanks (Doris; Ecuadorian, resident over four years; 2005)

Hello […] if you are going to get divorced I tell you the following, because it happened to me: 1) After the divorce they don't go chasing you to take away your residence permit, that is, the police. 2) You need to start the process [(applying for nationality)] before you start the divorce and then play around with the timings, I mean, hold on a bit longer. 3) I first applied for nationality and after a year I started the divorce process, and by the time of the interview with the police [(required for the nationality application)*] I had already started the divorce process. 4) My timings are as follows: Nov 2004 I start the process, mid 2005 interview with the police, divorce Jan 2006, nationality Dec 2006, I hope that this is useful to you, good luck.* (Andres, Colombian, 9 years of residence [original dates modified]; 2007)

Discussion of results

The importance of Spanish ancestors in shaping the patterns of access to Spanish nationality is a key finding of this analysis. This seems to be a fundamental factor in explaining the asymmetric relationship between, on the one hand, the frequencies of those internet forum users interested in acquisition of nationality by country of origin, and on the other, the actual distribution in official statistics of total immigrant population by country of origin, led by Romania, Morocco, Ecuador, United Kingdom, Colombia, and Bolivia (Instituto Nacional de Estadística, 2012). Moreover, another explanatory factor in this asymmetry is the unequal level of access to the Internet and general lack of 'web literacy' of persons from some of these countries, both in the origin country and in Spain, an aspect that is also related to difference in social class between migrant groups. This could be decisive in explaining the low volume of internet forum activity observed for people with origins in Ecuador and Bolivia.

From this analysis we can also conclude that there are three broad mechanisms that facilitate access to a European nationality; a) transgenerational migration through the *ius sanginis* provisions of some countries, key for destinations of historic European emigration; b) preference given to countries of the former colonial sphere of influence, and; c) the history of recent immigration flows that generate family re-unification movements, mixed marriages and transmission of the newly acquired nationalities to descendants regardless of the geography of residence.

Furthermore, beyond the Spanish nationality this analysis has also shown the long-term

consequences of historic emigration flows from other European nations to Latin America. Today, these flows have facilitated what is known as 'three-way migration'(Durand & Massey, 2010), in the study presented here primarily represented by Argentineans and Uruguayans with Italian or German origins that 'recover' the nationality of their ancestors but that prefer to live in Spain, because of linguistic and cultural proximity (Tintori, 2009). Furthermore, there are other cases of Latin Americans that after acquiring Spanish nationality through residence decide to live in other EU countries, such as Colombians in the UK (Guarnizo, 2008)

Now it is common to see Madrid full of Latin-Americans who are grandchildren of Italians, living legally thanks to their nationality being recognized, while the grandchildren of Spaniards are in many cases living illegally here. (Arturo, nationality unknown; 2004)

Another important aspect stemming from the analysis of this internet forum is that migrants elaborate strategies to access an EU nationality according to the most effective route for themselves and their families. Such strategies seek to maximize their ethnic capital within a system of hierarchical preferences in access to nationality established by each immigration country with respect to the rest of the world, as described in section 2. Therefore, migrants attempt to maximize their chances according to the most favorable combination of their personal circumstances with respect to three aspects; ancestors, marriage, and places and timing of residence. Within their personal combinations of these factors they develop a migration and nationality strategy through the country and legal mechanism (ancestors or residence) that presents the most favorable route in each case and point in time. In the context of a highly integrated European Economic Area (EEA), which of the 31 member countries is finally chosen is not that relevant. Instead, what migrants weigh is the number of bureaucratic hurdles and migratory requirements that need to be overcome in order to become an EU / EEA citizen following the shortest route for each individual's circumstances.

Furthermore, the data presented here and the individual cases highlighted in the text, support the hypothesis of a continuous weakening of the single or unique nationality as the fundamental model of belonging to a Nation State. Moreover, in the Spanish case, migrants who become naturalized also have to integrate into very different regional cultural and identity contexts, having to learn additional languages and customs, such as Galician, Basque or Catalan. Therefore multiple affiliations and spheres of identity are likely to be created and maintained, which involve the countries of origin, destination, 'three-way countries' as well as regional identities.

These findings also corroborate the general perception in the citizenship literature of the increased value placed by migrants on having multiple citizenships. The individual cases reported here show that even for U.S. citizens the possibility of possessing an EU passport is very attractive, in terms of work and residence possibilities across Europe, but also to enable access to welfare benefits that might not be available to them in the U.S.

In Latin America many of these combinations of pathways leading to multiple citizenships have been made possible by the historic confluence of migratory flows from a large range of different world regions. Having an ancestor from another country forms part of an 'ethnic capital' that is independent of, or complements, an individual's social capital, and becomes key in facilitating migratory flows. A person's surname, ethnic origin, phenotype and genealogy have recently become key components of migratory strategies at a global level. Regardless of whether one actually migrates or not, having the possibility of drawing upon one's 'ethnic capital' could be conceived as a life insurance policy, or a 'blank check', that is readily available when most required.

Future research avenues towards 'automatic netnography'

Stemming from this initial netnographic research there is a clear need to further automate the tasks involved in the different steps, as well as being able to develop techniques that work with different

languages. In Table 3 we identify five large internet discussion forums on migration and citizenship in Spanish and English that will be analysed in the next step of the project. The existing pre-classified Spanish discussion forum analysed in this paper, can be easily used as a gold standard to extract common expressions and validate the development of new automatic text-mining algorithms using natural language processing (NLP) techniques. These will be applied to the extraction of key information from the messages sent to these forums, attempting to identify pre-established variables as well as emerging themes. These could include; relevant information about countries involved, mobility trajectories, legal routes, ancestry information, intermediaries or service providers used should be automatically extracted and classified using a text-mining software application developed for this purpose. For example, the field of entity detection has made good progress in identifying and disambiguating known entities such as placenames (geocoded to a pair of geographical coordinates), people´s names, companies, and so on. Transdisciplinary research collaborations between social scientists and computer scientists working in text mining, semantic web, sentiment analysis, entity recognition, computer linguistics and so on. Apart from collecting standardised information, the automatic classification system should also collect relevant excerpts from the messages reflecting wider general user views and perceptions. As a result of this phase of automatic analyses of messages' text, users and discussion threads should be classified into broad types of migration trajectories and general topics of concerns; for example, ancestry, naturalisation, marriage, residence permit, irregularity, return, welfare, and so on. In a second phase, the messages from these typologies of users should be then manually coded using qualitative text analysis techniques and expert intervention. Manual analysis should attempt to identify in-depth the key drivers and motivations for choosing or rejecting access to multiple citizenship. These *a priori* could be a combination of factors such as: country of birth and of socialisation, countries of residence and nationality; life stage; ancestors' origins; ethnicity; language ability; social class; information about legal routes available; changes in origin/destination legislation or labour markets; family imitation; and expectations about the perceived value of EU citizenship. It is important to establish a feedback mechanism between this manual, second phase and the identification of potential intervening factors in migration and multiple citizenship, in order to identify key issues and themes to be fed back to the automatic search system as new key expressions. This cyclical process could constitute a virtuous loop in which the manual identification of such intervening factors in the forum discussions, leads to automatic finding of further cases under the same circumstances. This process could be refined through a calibration of search terms y the dictionary of expressions as well as of the search rules applied. Finally, futher meta-classification of particular users, through socio-demographic profiles could also be applied following crowd-sourcing non-expert classification of general user characteristics (McCormick et al., 2013).

Through the pilot netnographic study analysed in this paper, a set of avenues for future research have been identified. These can be grouped into four general areas, as summarized in Figure 3:

1. **Data harvesting**; involves issues of automating the actual data harvesting from the web, navigating through the website, establishing initial parsing around HTML tags in each webpage, and extracting and storing the text retrieved into an structured database. This aspect also requires being able to adapt the harvesting engine to the various platforms and standards followed by each internet discussion forum. All forums identified in Table 3 run in software platforms and HTML in English.
2. **Text mining**; The set of text mining techniques need to be replicated and adapted for each language (in this case English and Spanish), with the ultimate aim to develop multilingual capability in the tools developed. A range of techniques borrowed from the field of Natural

Language Processing should be applied. These include identifying sentences, where possible, and use a parsing algorithm to breakdown its grammatical structure identifying the subject and the object of the sentence. There are number of parsing algorithms available in English and Spanish, the use Stanford parsing algorithm being commonly used in English (Marneffe & Manning, 2008). However, this can be inefficient when analysing thousands of web messages, especially since a large proportion of social media writing is not always grammatically correct. Therefore, other approaches called 'shallow parsing' and 'parts of speech' (POS) analysis are required to identify different components in a sentence such as verbs, nouns, adjectives/adverbs, prepositions etc. These structurally stored text can be then mined for relationships between common expressions (Black, Procter, Gray, & Ananiadou, 2010). Using a set of rules custom-built for the topic of migration and citizenship, users in the internet forum will be classified according to the variables identified earlier in this section. Furthermore, common expressions will be searched for in the entire database using multiple word frequency engines, such as *Termine* tool developed by the National Centre for Text Mining at the University of Manchester (Frantzi, Ananiadou, & Mima, 2000). With these a dictionary of expressions will be built, each referenced to a standard topic of interest in migration and citizenship (e.g. naturalisation, ancestry and so on)

3. **Data Analysis,** involves a further phase in which a set of flexible queries and language relationship rules will be built in order to retrieve relevant excerpts of text and classify user profiles according to the aforementioned trajectories and typologies. In this phase, new key topics suggestions will be arising from the text (emic codes). Further semi-automatic analysis will be applied to establish temporal trends and relationships between user types.
4. **Ethics**. This type of netnographic research brings new ethical considerations that have not been discussed in this paper but will need to be properly addressed by new research in this area. What are the ethical implications of monitoring public opinions on the social web? Is there a need for informed consent from the passive participants in the study? British academics have formed a network of researchers termed 'New Social Media New Social Science' (http://nsmnss.blogspot.co.uk/) that is already dealing with these ethical issues and setting recommendations for future research conduct.

Conclusion

Netnographic research methods in the field of migration and citizenship have been evaluated in this paper through a pilot project. This project comprised automatic and manual phases of analysis of a large body of text derived from an internet discussion forum on access to Spanish citizenship. Netnography, as an innovative research method, presents a series of advantages and some weaknesses. Amongst the former, it offers a very economical alternative to traditional survey and interview methods of data collection, it includes very large samples spread over large geographical distances and temporal periods (retrospective analysis is feasible), there is no cost of transcription of audio files, and in general it provides honest and detailed responses from discussion forum participants. Amongst the weaknesses, it necessarily provides incomplete life histories from research subjects, and the whole research design lacks the control of the typical social science project; i.e., questionnaire/schedule design, sampling, quality filtering, and measuring margin of error and representativeness. However, as demonstrated by the value of the results coming out of the pilot project presented in this paper, its benefits clearly outweighs its disadvantages, especially when compared to alternative more traditional methods to study complex and rapidly changing international migration and citizenship practices.

New research in this area should attempt to automate the tasks of user profile classification and topic detection. Some clues for future research applying text mining and natural language processing algorithms to social media content have been outlined, configuring the basis for future attempts at taming the mass of big data produced by millions of individuals interacting in social media technologies. The findings and recommendations posed in this paper for internet discussion forums can be extrapolated to other social media platforms. These findings point towards very promising methodological developments for the social sciences over the coming decade. Population scholars in various fields are encouraged forge new trans-disciplinar collaborations with computer and physical scientists, at the same time as bridging the gap between scholars in the quantitative and qualitative traditions, in order to embracing this innovative research methodology. This approach will almost certainly offer them unexpected insights in understanding population change in a rapidly moving world which traditional demographic research methods, especially in migration, are struggling to monitor.

References

Black, W., Procter, R., Gray, S., & Ananiadou, S. (2010). A data and analysis resource for an experiment in text mining a collection of micro-blogs on a political topic, 2083–2088.

Durand, J., & Massey, D. S. (2010). The ANNALS of the American Academy of Political and Social Science Orders : doi:10.1177/0002716210368102

Frantzi, K., Ananiadou, S., & Mima, H. (2000). Automatic recognition of multi-word terms:. the C-value/NC-value method. *International Journal on Digital Libraries*.

Golder, S., & Macy, M. (2012). Social Science with Social Media. *Footnotes*, *40*(1), 1–20.

Guarnizo, L. E. (2008). *Londres latina: la presencia colombiana en la capital británica*. Zacatecas, Mexico: Universidad de Zacatecas / Porrua.

Hesse-Biber, S., & Leavy, P. (2008). *Handbook of Emergent Methods*. London: Guilford Publications.

Instituto Nacional de Estadística. (2012). *Cifras oficiales de población*. Madrid.

King, G. (2011). Ensuring the data-rich future of the social sciences. *Science*, *331*(6018), 719–21. doi:10.1126/science.1197872

Kozinets, R. V. (2010). *Netnography: Doing Ethnographic Research Online* (p. 232). London: Sage Publications Ltd.

Marneffe, M. De, & Manning, C. D. (2008). Stanford typed dependencies manual, (September), 1–19.

McCormick, T., Lee, H., Cesare, N., & Shojaie, A. (2013). Using Twitter for Demographic and Social Science Research: Tools for Data Collection.

Miller, D. (2011). *Tales from Facebook* (p. 220). Polity Press.

The World Bank. (2011). *Atlas of Global Development*. Washington D.C.: World Bank Publications.

Tintori, G. (2009). *Fardelli d'Italia? Conseguenze nazionali e transnazionali delle politiche di cittadinanza italiane* (p. 127). Carocci.

Tables and figures

Nr.	Variable name	Pre-set responses	Description
1	ID	Person ID number	ID number internally asigned to each user
2	Role in the group	Participant / Leader	Main role of the user in the forum: participant, who only enquires about his/her own case, or leader, who responds to others' enquiries
3	Country of origin	Country name / n.a.	User's country of origin or former country of nationality
4	Spanish Ancestors?	Y/N / n.a.	The user [has/ does not have / we do not know] Spanish ancestors
5	Seeks nationality	Y/N / n.a.	Is the user explicitly interested in applying for Spanish nationality?
6	Year of nationality application	Year / n.a.	(If answered Y to question 5) year of initial application of nationality
7	Nationality approved?	Y/N	(If answered Y to question 5) Has the nationality application been succesful?
8	Year nationality acquired	Year / n.a.	(If answered Y to question 7) Year of nationality acquisition
9	Method of access to nationality	residence / spouse/ parents/ grandparents/ off-spring/	Main method to access Spanish nationality according to the legislation at the time of participation in the forum.
10	Resides in Spain?	Y/N / n.a.	Does the user live or has lived in Spain in the past as a resident?
11	Year of arrival to Spain	Year / n.a.	Year of first arrival to Spain
12	Length of residence in Spain	< 1 year / 1-2 years / 3-5 years / > 5 years / n.a.	Length of residence in Spain (either derived from question 11 or taken from the text)
13	Signs of irregular experience?	Y/N/ probably	Does the user present any signs of irregularity in his/her migration legal status (either present or past status/es)
14	Undocumented/irregular migrant?	Y/N / n.a.	Does the user explicitly report to have been in irregular migration status?
15	Returned to country of origin?	Y/N/ 'intermitently' / intends to return	Has the user returned or intends to return to his/her country of origin? ('intermitently' means any type of circular movement)
16	Other observations	Free text	Any other observation that the research assistant deems to be useful for the investigation, such as flagging up relevant quotes for qualitative analysis

Table 1: Description of variables captured for each forum's user in the research database
Source: Author's database compiled from Registro Civil internet forum

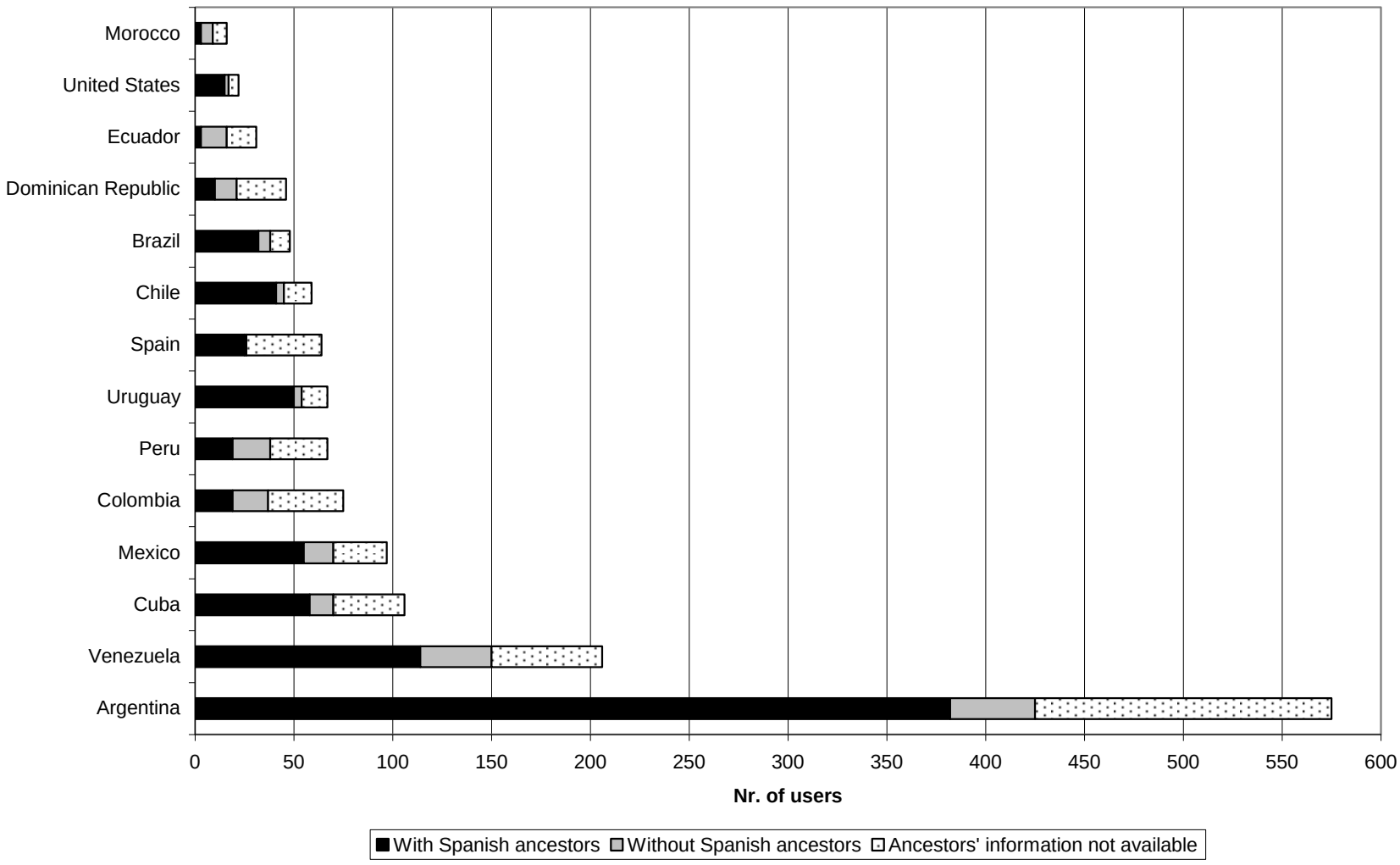

Figure 1: Frequency of users by country of origin and presence of Spanish ancestors (top 14 countries)
Source: Author's database compiled from Registro Civil internet forum

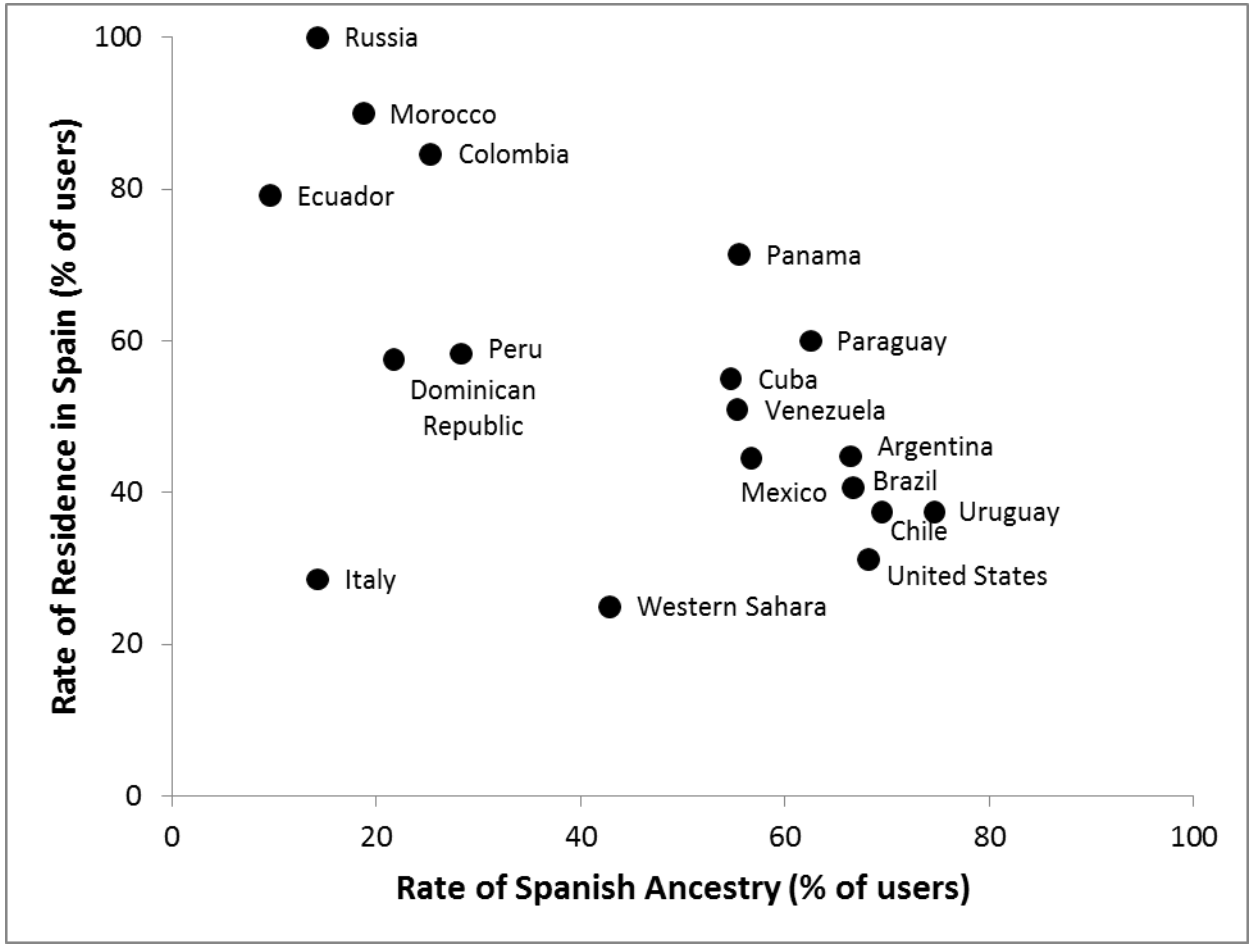

Figure 2: Scatter plot of rate of Spanish residence vs. rate of Spanish ancestry by country of origin

Rate of Spanish residence is the percentage of users by country of origin who declare to have ever resided in Spain, while the rate of Spanish ancestry is the percentage of users by country of origin who declare to have Spanish ancestors. Only countries with a least 7 users for both types of rates are shown.
Source: Author's database compiled from Registro Civil internet forum

Country of origin (1)	Total People (2)	%Ancestry (3)	%Spouse & Children (4)	Difference quotient (5)	Predominant Method (6)
Ukraine	5	0%	80%	1	
Russia	6	0%	33%	1	
Morocco	10	10%	60%	0.71	
Dominican Republic	37	16%	65%	0.6	**Spouse & Children**
Peru	53	17%	49%	0.49	
Colombia	61	16%	38%	0.39	
Ecuador	23	22%	43%	0.33	
Panama	8	25%	38%	0.2	
Paraguay	8	25%	38%	0.2	
Portugal	5	20%	20%	0	**Neutral**
Mexico	80	39%	25%	-0.22	
Venezuela	165	53%	28%	-0.3	
United States	16	50%	25%	-0.33	
Cuba	78	45%	18%	-0.43	
Argentina	469	39%	15%	-0.44	
Brazil	36	47%	14%	-0.55	**Ancestry**
Uruguay	63	43%	11%	-0.59	
Chile	41	49%	12%	-0.6	
Western Sahara	6	33%	0%	-1	

Table 2: Countries of origin by predominant type of 'family route' to access to nationality

Countries of origin included (column 1) are those with at least five users with a valid response in variable 9 (method of access to nationality), comprising the total number of users per country (column 2) to which the rest of percentages refer to. The next two columns report the percentage of users that apply for nationality through: ancestors (column 3) and spouse or children (column 4). Column 5 shows a difference quotient between the previous two columns (see text, equation 1), while column 6 classifies the 19 countries into three predominant types of family route to access nationality.

Source: Author's database compiled from Registro Civil internet forum

Discussion Forum Name	***Dates***	***Nr. of messages***	***Link***
Spanish Civil Registry forum 1 (Historic site)	2001-2008	54,920 messages (16,856 on nationality) / 6,813 people	*Site discontinued*
Spanish Civil Registry forum 2 (current site)	2009-	22,000 messages (17,300 on nationality)	http://www.fororegistrocivil.es/Foro/viewforum.php?f=4
Exiliados, Spanish citiz.	2008-	5206 messages / 4612 people	http://exiliados.org/foro/
British Nationality - UK Citizenship	2008-	19,500 messages (all on nationality)	http://www.ukresident.com/forums/forum/2495-british-nationality-uk-citizenship-and-naturalisation/
Immigration Boards – British Citizenship	2002-	35,253 messages (all on nationality)	http://www.immigrationboards.com/viewforum.php?f=42

Table 3: Examples of internet discussion forums on migration and citizenship subject to netnographic analysis

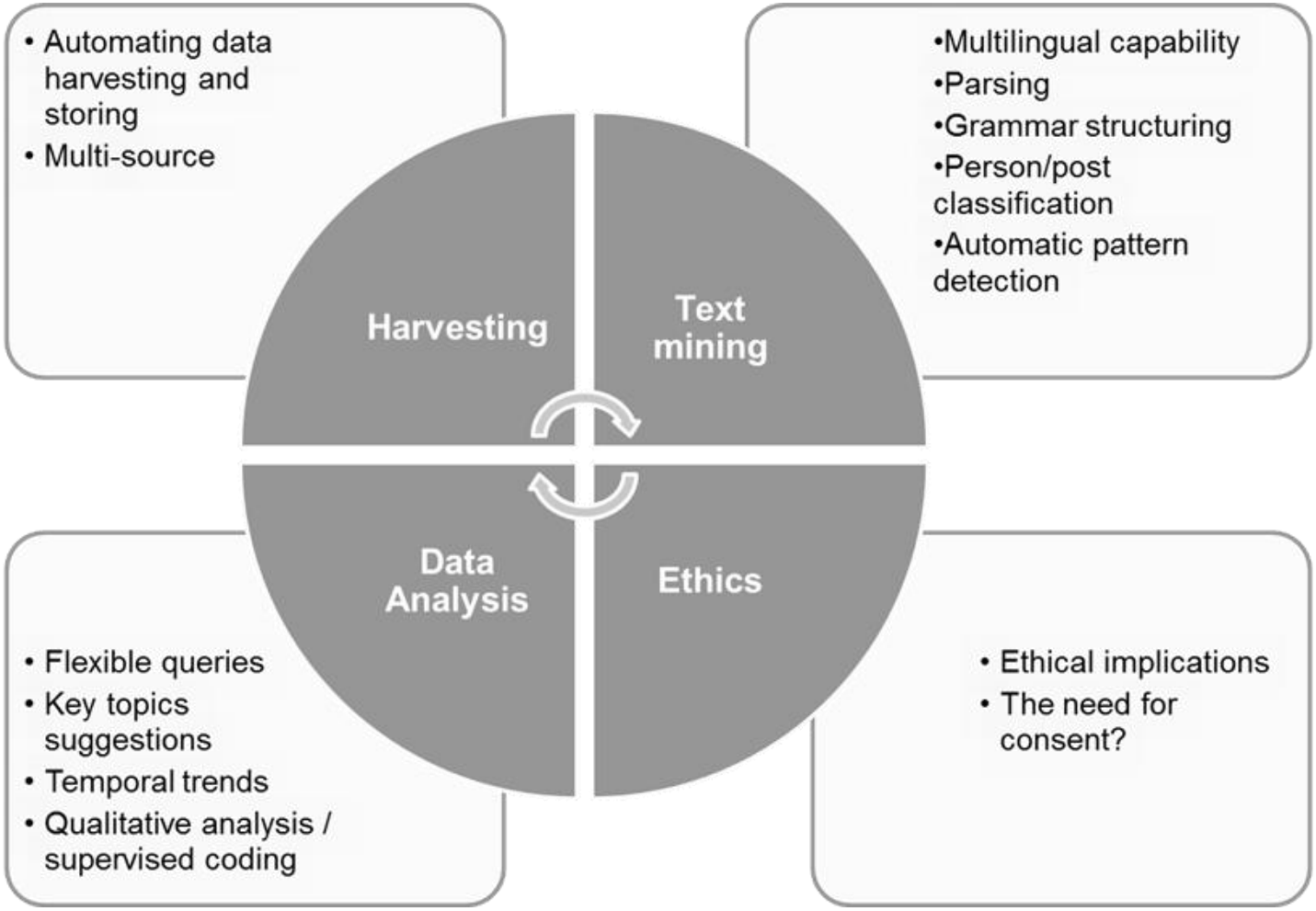

Figure 3: Summary of research challenges in future *netnographic* analysis of internet discussion forums