Analysing User Generated Content for social science. Generational "we sense" in the Italian blogosphere.

Prof. Giovanni Boccia Artieri[1], Dott. Luca Rossi[2], Dott. Fabio Giglietto[3]

[1, 2, 3] Advanced Communication Research Laboratory, University of Urbino "Carlo Bo", Italy

[1,2,3,] giovanni.bocciaartieri / luca.rossi / fabio.giglietto @uniurb.it

Abstract. The goal of this paper is to present an innovative methodology to exploit user generated content as a data source for sociological research. The methodology will be presented by discussing a specific research case study project. The discussed research project goal is to describe the role of media contents in the construction of generational identity through a two step question. May specific media-products get user generated generational discourses started? If so, may those discourses be used to investigate the shared generational *we sense*?

Web 2.0: what has been changing for social sciences

During the last few years the Internet has been increasingly used by people as a read-write medium. Thanks to the dropped prices and skills necessary to afford and use technologies to create digital contents (such as camera phones, video and digital cameras), people are now able to create persistent digital information. A large share of this information is today exposed to a mass audience on the Internet. For the first time in history people have access to mass medium not just on audience side but also on the producers' one, and users generated contents (UCG) have been emerging everywhere on the net. Worldwide, there are over 70 millions of tracked weblogs[1], more than 500 millions of photos[2] and 240 millions of videos[3] available on the Internet. The actual grow rate of these user generated contents is amazing and unprecedented. We are in the era of "mass communication for the masses". This amazing change in the way people is using Internet is affecting many countries all over the world. The current Italian scenario, which has traditionally been affected by an adoption gap, also appears to be encouraging. According to a recent survey (LaRiCA 2008) the 6% of Italian over 18 population is a weblog's author and the 9.6% has a profile on a social network sites of any kind (Myspace, Facebook, Badoo, etc.). Beside that growing production of online contents there is a parallel growth of content sharing. The previously reported research states that the 12.1% of Italian over 18 population usually shares online pictures (the percentage drops to 9.5% for video sharing, to 9.1% for text sharing and to 5% for audio sharing). In

---

[1] David Sifry, The State of the Live Web, April 2007, http://www.sifry.com/alerts/archives/000493.html.

[2] Lele Dainesi, Flickr sarà in italiano da stasera alle 18:00, http://www.leledainesi.com/archives/2007/06/12/flickr-sar-in-italiano-da-stasera-alle-1800/ (Flickr press conference data).

[3] Google video data (also counting YouTube) retrieved on May 26, 2008.

addition to that the very low percentage of people who declare being worried about the copyright of the contents they share (28%) may lead to the idea that this shift toward a *production and sharing* approach to the web is deeply rooted into online users, despite every legal attempt that has been made to restrain it.

This growing amount of online data, i.e. blogs, has four characteristics that tend to increase even more the sociological value of these conversations. Boyd (2004) has suggested that the online network of communications is persistent, searchable, replicable and addressed to an invisible audience. Each of these properties is nothing new in the media landscape but all of them have never been present in one media before the World Wide Web.

a. Persistence is one aspect that online communications share with other media (e.g. writings). It is possible to access online communication even after the communication event has taken place and sometimes link structures describe the process of reproduction of communication from communication. This point is useful for longitudinal research.

b. A large amount of data is not always useful if there is no way to filter and sort the data according to research questions. Online conversations as digital texts are easy to search (e.g. Google). In addition, digital photos and videos are often also easy to search thanks to the tags that users use to apply in order to categorize their own contents. At the same time the structure of tags becomes an observable item and can be even used as data (e.g. social bookmarking tools such as del.icio.us or Flickr tag clouds).

c. Online conversations are easy to replicate. This property increases the possibility of re-using and mixing of available contents and creating the possibilities to observe, along with the quotation link structure, the process of reproduction of communications.

d. Online communication can be addressed to an unknowable audience. In the absence of a direct way to identify audiences, the author of a blog can create their own image of their potential readers and choose the properties of the page (e.g. themes, style, etc). On the Internet everyone can be an author and expose its contents to the masses.

The contemporary presence of those characteristics make the autonomously produced data already available on the Internet, a valuable resource for every social researcher. These properties, especially the fact that online conversations are searchable, allow unfolding a whole new set of research possibilities. Traditional researches on online conversations have mainly been based on Virtual-Ethnography that requires the "situated presence of an ethnographer in the field setting, combined with intensive engagement with everyday life of the inhabitants of the field site" (Hine 2000, p. 63). This strategy offers to the researchers a deep qualitative understanding of online phenomena but at the same time it tends to restrict the numbers of conversations which can be studied. At the same time, while an ethnographic approach makes easy to observe a specific online place (e.g a single newsgroup, a chat room, a single virtual world, etc.) it seems to have difficulty when there is a need to observe how a specific discussion develops on the Internet. Generational issues are a perfect example of a Internet-wide topic which can be discussed almost everywhere on the net.

Media and Generations: theoretical perspective

The main goal of the research project we are discussing here is to understand if and how media products (e.g. novels, movies, TV shows) affect the wide process leading to the creation of a shared set of meanings and sense of belonging, what Bude (1997) called generational 'we sense'. According to Mannheim's classical definition of the sociological concept of generations (Mannheim 1952), a generation arises when the youth experience the same concrete historical problems. This classical definition, which in a direct way links the construction of a set of shared meanings with historical problems, even if it provides a fruitful starting point for every further analysis, it doesn't explain much about how generations arise. It has been argued that people who grew up together and faced the same historical problems can be as well defined by using the "technical" concept of 'cohort' (Ryder 1965, Glenn 1977). The cohort concept keeps together people according to their birth year and with regard to the characteristics they share (Corsten 1999). Compared to the concept of generation, the cohort one seems to be much more easy to define and it allows a large comparison between different birth cohorts. In spite of the success of the cohort concept the generational concept seems still to be attractive for sociologists. The last theorizations about generations try to explain the concept not as a mere construction operated by the sociologist (Corsten 1999, Edmunds 2002), but as a multi-dimensional issue which assumes a shared assumption of a common life experience by the members and the emergence of that specific generation as a social-fact itself. The generation, as a theoretical concept, then assumes a reflexivity process (Giddens 1991) both at an individual and at a collective level (Giddens, Beck, Lash 1994). Recently Edmunds and Turner (2005) proposed that mass media might be a common landscape able to offer a world-wide shared stage for "concrete historical problems" experienced by a growing *global generation*. The global media-scape (Appadurai 1996) offers the possibility for young people all over the word to experience, for the first time in history, a global view of the world and, as a consequence, a global dimension of problems. From the Vietnam war to 9/11 several waves of global, media-based, generations might experience specific localized problems as world-wide issues. If the generational 'we sense' may be described from a more hermeneutical and linguistic point of view as a meaningful set of connected criteria for interpreting and articulating topics in communication (Corsten 1999, Luhmann 1980), then mass media would be the place where those 'criteria' are learnt. Edmunds and Turner's assumption about a global generation (2005) moves a step forward the whole generational research. If, as we claimed before, generations are a shared construction that cannot exist without a specific generational perception between the members, how can this "we sense" in world-wide media based generations be observed?

Weblogs, and the whole web 2.0, seems to be a viable place where to observe if and how generational discourses emerge, and if they may arise around specific media-contents or media-topic able to trigger the reflexive process.

WeSearch methodology

In order to investigate media-related generational discourses on the web we have developed a research process called 'WeSearch'. It has been articulated in four steps: a) *identification of generational products*, b) *definition of query strings for online analysis*, c) *retrieval of blogs author's information d) content analysis of collected data*.

Identification of generational products

The research process started by identifying a set of media products (movies, music albums, TV-shows, books and comics) which had been labeled as generational by their audience[4]. In order to do so we carried on 5 group-interview with 3 participants belonging to generations X and Y, who were born between 1966 (the beginning of Generation X) and 1991 (the last birth year of generation Y). That specific generational time-frame had been chosen according to an existing literature (Aroldi, Colombo 2007). Each group had been asked by the interviewer to identify a set of media products that they would consider "generational". During the interviews the researchers focused the attention not only on media products but also on how media had been experienced and how media-products entered the generational discourse. This phase ended with a set of 45 media products (chosen from the top 9 of each category).

Definition of query strings for online research

In order to collect all the Italian-language blog posts dealing with selected media products, every media-product had been used as a keyword to query Google Blog Search. Even if we were aware of some of the weakness of Google Blog Search for social analysis (see Escher 2008), we decided to use Google's service in order to collect blogs entries mainly for two reasons: 1. it returned less SPAM blog posts than similar services and 2. the language filter worked better than competitors such as Technorati, allowing us to focus the research only on Italian blogs. Google blog search had been queried in order to obtain an RSS feed containing top 100 entries ordered by relevance. We opted for a relevance sorting because of the non chronological nature of collected data. The outcome of this phase was a standard RSS feed generated by Google's service, composed by every retrieved blog entry containing selected keywords. The first problem that the research team encountered was that sometimes blog services did not provide an RSS feed containing the full text of the blog entry. To solve this problem it was necessary to develop a tool able to fetch the full text content of every single article. We chose to use the 'Yahoo pipes' services in order to develop a small tool aimed to carry on this work for us. The final output of this phase is an RSS feed containing the full text of every single blog entry containing the selected keyword.

Storage of blog entries and retrieval of authors' information

The RSS feeds was stored in an incremental way using a software which was specifically envisioned by the research team. The web application is able to store an RSS feed, sorting the stored information by authors. Another feature of this software is the ability to retrieve biographical information about entries authors. Obviously this information may be obtained only if the author him/herself added them to his/her own online profile. Biographical information (age, gender and location) is retrieved using a scraping technique specifically developed for every blogging or sharing service. Currently there are 7 supported platforms, as shown in table 1.

---

[4] The research team didn't provide a definition for the term "generational" but let the participants interpret the term according to their own perception.

service	scraping languages	url	unique audience [000][5]
Blogger	ITA / ENG	http://www.blogger.com	5190
Flickr	ENG	http://www.flickr.com	N/A
Il Cannocchiale	ENG	http://www.ilcannocchiale.it	590
Libero	ITA	http://www.libero.it	4643
Splinder	ITA	http://www.splinder.com	2303
Windows Live Space	ENG	http://spaces.live.com	5063
YouTube	ENG	http://www.youtube.com	N/A

Table 1

All the collected information (both the posts and the biographical data about authors) may, at any time, be exported in textual files (.doc format). The exported files will be compatible with the Nvivo7 qualitative analysis software. Biographical information about authors will be exported in a comma separated data file which can be easily imported as an Nvivo casebook retaining the link between the biographical information and the text entries.

Summary of collected data

Starting from the set of 45 media products used as starting keywords, the research was able to retrieve around 3000 blogs entries. Author's information was available for 928 cases (31%) of which 49.34% was male and 50.76% female. Generation Y is the most represented with the 79% of blogs entries[6], generation X scores the 15% and Boomer Generation (born between 1953 and 1965) and postwar (born between 1940 and 1952) score both 3%. The collected data were analyzed from the qualitative and quantitative point of view using Nvivo7 software.

Before stepping into the research results and move back to the relation between media products and generational we sense, it is now possible to make few preliminary considerations on the adopted methodology. One of the most interesting things we noticed is somewhat a confirmation of what it is already well known about American blogosphere (Lenhart and Madden, 2007). Italian bloggers use to share thoughts and personal information (e.g. age, gender on profile) online in a massive way. The scraping technique we developed (that can be easily extended thanks to its modular structure) was able to collect authors' information for more than 1/3 of the whole set of data (31%). The data confirm that the vast majority of active Internet users (blog authors, people who share pictures or video online, etc.) may be located in the youngest part of the population. According to the previously mentioned research (LaRiCA 2008) the 74.3% of Italian blogs authors is between 18 and 29 years old. Overall results are, of course, heavy influenced by this age distribution.

Together with those encouraging results we also experienced some limits of the methodology. The first obvious objection concern the author's biographical data in the profiles. We are well aware that data in user profiles may not be accurate since the users may decide intentionally

---

[5] Data courtesy of Nielsen//Netratings Netview. Standard Metrics (Internet Applications Included) February, 2008, Country: Italy. YouTube and Flickr was not considered .

[6] Obviously all the information about generations is based on the smaller set of blogs entries with author's information available.

to deceive while filling the registration form. At the same time we also know that both post contents and data in the profile are conversations and we therefore observe them as a social construction. On this level of observation what matters is the realness of the construction itself and not its relationship with reality (Luhmann 1980).

Does the imaginary status of the characters in a novel affect the sociological value of the content? Our study on weblogs may be compared from this point of view to a new type of sociology of literature with the big difference that today everyone may take part to the public collective narration of our society.

A more tricky limit to overcome is the difference between the age as reported by the author profile and the date the content was produced. Even if the platform translates the author's birth date in its present age we still cannot be sure about the age of the author when she/he produced the content[7].

On a completely different level we should also note that the amount of data retrieved requires a large effort to be qualitative analyzed. The application we chose to carry on this part of the job[8] unlashed its limits both on the performance side and on the workgroup features.

At the same time the filters we created, even if accurately tested, also retrieved a large amount of unrelated and useless content. A better filter based on words correlation and maybe semantic awareness would definitely decrease the time required by the content analysis.

Posting about Generational Identity[9]

The huge number of retrieved posts pinpointed the capability of media products to trigger generational discourses. Generational discourses, triggered by the specific media product, seem to be used in two different ways. Sometimes the specific product gives the opportunity to start a reflexive process narrating single and personal events. Media product may then act as specific keys to release hidden memories or as pivotal elements in the personal biography.

> "I chose to start the nursery class. My nurse-syndrome and the fact that I watched *Candy Candy* when I was a child, made me choose this university degree" (F., female, 19 years old)

Media-products are perceived as something deeply related to the individual life and identity creation. Despite that, at the same time, media products seem able to start wider reflexivity processes. Those processes, which may be considered as a truly generational form of thought, seem to link the specific media product to a wider and shared *we sense*.

> "Today there are no more values in society… but today's television is partially guilty… do you remember the cartoons we used to watch and how they taught us to love and take care of feelings?! Candy Candy, Lady Oscar and Georgie… What today's cartoons are about is only violence!!!!" (F., female, 25 years old)

---

[7] A possible way to overcome this limit would be to calculate the age of the author using both the age on profile, the publishing date of the content and the date the content was retrieved by the researcher.

[8] QSR Nvivo 7.

[9] All the blog excerpts were translated from Italian to English by the researchers.

There is the feeling of something shared. Something that is assumed to be common because of the sharing of a specific media product or of a specific time in media history:

> "And we who are in the age of thirties, we belong to the Tiger Mask generation, we can't change it […] We had good times, watching cartoons and TV series that left their mark." (A., male, 32 years old)

This kind of double reflexivity that media products allow (individual reflexivity and generational reflexivity) is strongly related to the Italian media history and to the heavy diffusion during the seventies and the eighties of media technologies. Media technologies, due to the high impact they have in everyday life-routine may be seen as milestones used to label specific time in personal and generational history. Generation definitions, when observed from the media perspective, seem to be a mixture of specific media-products and specific media-technologies:

> "We the generation born in the seventies and grew up in the eighties […] we're the last *naïve* generation. We didn't have super videogames and virtual reality, our games were the crazy ball, and the gluing hand… the bicycle (Graziella [a folding bicycle] for girls and BMX for boys). The top edge technology we had was Pac-Man" (E., male, 30 years old)

Conclusions

This paper aimed at the same time to show a new methodology for using UGC in social sciences and to test the proposed methodology on a research about the relationship between media products and generational perception. From the methodological point of view this paper is a contribution to the ongoing debate about how to use the so called web 2.0 technologies and practices for sociological research (Beer, Burrows 2007). The conversational form of the contemporary world wide web, the increasing availability of spontaneous UGC and the possibility to search and process huge amounts of such contents, make the social web a very interesting resource for every topic of social research. The proposed methodology, that we called *WeSearch*, is a mixed process which freely uses available online tools (such as Google blog search and Yahoo pipes) as well as specifically designed software. The most interesting aspect of the *WeSearch* methodology is the capability to retrieve a large amount of spontaneous qualitative data everywhere on the Internet. Such a methodology combines the high value of quantitative research with the in-depth analysis typical of a qualitative approach.

The reported case study shows some interesting results. The good quantitative result of the analysis demonstrates that media-products are really good triggers for generational discourses. It is interesting that the specific-media product is not necessarily the topic of the produced content. People don't usually speak about that media-product but they use it as a specific mnemonic anchor. A single product may be used to help the reader remember what that time was or what the general "mood of the time was". Truly generational media-products do not require introductions or explanations since the writer assumes that readers know them. The mutual knowledge is taken for granted. F., in our first example, does not care to explain what the relationship is between *Candy Candy* and nursery since she assumes that everybody should know it. If media-products seem to be very good anchors for memories, the generational use of these memories seems to be a little bit more complex. Authors usually talk about media contents both in an individual and in a generational way. We consider an individual use of the media-product when the product is linked to a specific event of the

user's personal life. At the same time a truly generational use of the media-product may be observed when the product is used to evoke a shared knowledge or feeling. In this second case the media-product acts as a strongbox for a shared *we sense* which may be summoned by the media-product. In the reported example A. speaks about *Tiger Mask* in order to disclose a shared knowledge not related to the anime itself but about the generation of the viewers. It is like a secret phrase known only to some.

If media-products showed to be a very valuable key in order to find generational discourses on the net, the research failed to obtain a full reconstruction of the generational *we sense.* An ambitious goal of the research was to be able to describe a different generational map of shared values starting from the retrieved online conversations. The main problem we encountered was that not every content may be easily be used to describe values. The produced contents often deal with personal life or general/generational belonging without showing a direct link with values. This made the generational semantic almost impossible to retrieve from data. If media-products seem to be very good in re-activating generational memories and the mood of the time, it has still to be understood how those ones might be used to investigate the specific generational values-map.

References

Appadurai, A. (1996). Modernity at Large: Cultural Dimensions of Globalization. Public worlds.. University of Minnesota Press, Minneapolis

Aroldi, P. and Colombo F. (2007): "Generational belonging and mediascape." Journal of Social Science Education. n.1

Bech, U. Lash S. and Giddens A. (1994): Reflexive Modernization: Politics, Tradition, and Aesthetics in the Modern Social Order, Stanford University Press, Stanford Calif.

Beer D. and Burrows R., (2007): "Sociology and, of and in Web 2.0: Some Initial Considerations.", vol. 12, issue 5, http://www.socresonline.org.uk/12/5/17.html

Boyd, D. (2007) "Why Youth (Heart) Social Network Sites: The Role of Networked Publics." in D. Buckingham (eds.): *Teenage Social Life. Youth, Identity and Digital Media*. MIT Press, Cambridge, MA, pp. 119-142.

Bude, H. (1997). Das Altern einer Generation. Die Jahrgange 1938-1948. Suhrkamp, Frankfurt am Main.

Corsten, M. (1999). "The time of Generations." TIME & SOCIETY, vol. 8, no. 2, 1999, pp. 249-272

Edmunds. J. (2002). Generations, Culture And Society, Open University Press, Philadelphia.

Edmunds J., Turner B. (2005), "Global Generations: Social Change in the Twentieth Century", The British Journal of Sociology, 56, 4 (pp. 559-577)

Escher, T. (2008). "Five lessons on how Google Blogsearch works (or doesn't) and how to use it for research." Tobias Escher at the OII. http://people.oii.ox.ac.uk/escher/2008/02/28/google-blogsearch-howto/.

Giddens, A. (1991). Modernity and Self-Identity: Self and Society in the Late Modern Age. Stanford University Press, Stanford, Calif.

Glenn, N. D. (1977). Cohort Analysis, SAGE, Beverly Hills, CA.

Hine, C. (2000). Virtual Ethnography. Sage Publications Ltd. London.

LaRiCA. (2008). I social media in Italia: oltre la punta dell'iceberg. http://larica-virtual.soc.uniurb.it/?p=39.

Lenhart A., Madden M. (2007), "Teens, Privacy & Online Social Networks," Pew Internet & American Life Project, Washington 2007.

Luhmann, N. (1980). Gesellschaftsstruktur Und Semantik: Studien Zur Wissenssoziologie Der Modernen Gesellschaft. Suhrkamp, Frankfurt am Main.

Mannheim, K. (1952). Essays on the Sociology of Knowledge. Routledge & K. Paul. London.

Ryder, N. (1965). The Cohort ad a Concept in the Study of Social Change, American Sociological Review, 30:843-61