Managing and Using Unstable Data in a Social Science Research about Museums and Audiences on Social Media

Marie Van Cranenbroeck
Université catholique de Louvain
Ruelle de la Lanterne magique 14,
bte L2.03.02
B-1348 Louvain-la-Neuve
marie.vancranenbroeck@uclouvain.be

ABSTRACT

In this paper, we will discuss the technical and methodological difficulties that we encountered with the social media data collection and the content analysis of these data during our doctoral research. More generally, we will examine some problems regarding the content analysis of big data within the scope of a social science research.

Keywords

Audiences, Big Data, Communication, Data Collection, Data Archive, Museums, Social Media, Social Science Research.

1. INTRODUCTION

This paper is divided in three parts. In the first part, we will briefly introduce the topic of our doctoral research and our research questions, and talk about our methodological framework. This part will allow us to answer the following question: 'What did we try to achieve?'. The second part will enable us to discuss the principal difficulties that we encountered during the data collection, the content analysis and more generally the problems of doing a content analysis of big data within the scope of a social science research. This second part will answer the questions: 'What went wrong or did not work as expected?' and 'What can others learn from this approach?' In the final part, as a conclusion, we will summarize several lessons that we have learned, starting from the specificities of our research, in an attempt to point out a few transversal questions that can be useful in other social science research studies which use big data.

2. RESEARCH QUESTIONS AND METHODOLOGICAL FRAMEWORK

The relationship between museums and audiences in the Internet era is at the base of our doctoral research. We wanted to determine and question this connexion, the 'clash' (or the absence of it) of the museum missions [1] [2] [3], the hopes and limits of the Internet and audience's expectations [4] [5] [6] [7].

For many years and despite the efforts of the museums, visitor surveys show that museums are seen as important places but also as rather elitist institutions. On one hand, the relationship between museums and audiences is mainly based on access (to museums, exhibitions, activities, heritages, knowledge, etc.). On the other hand, the presence of the museums on the Internet and social media ignites again the hope of a more balanced relationship, based on the interaction and maybe even the participation of visitors and users. We found that Nico Carpentier's AIP model [8] [9] which makes a distinction between access from interaction and participation was particularly interesting to enlighten the relationships between users on the Internet… and, of course, between museums and audiences. Is this relationship more balanced on social media? How are museums using social media? Do audiences really want to interact with museums or participate in the digital museum life?

In order to answer these questions, our research is based on three stages which are independent of each other and are not hierarchical: an *in situ* and online survey which reached 1000 respondents, 40 interviews with museum teams and visitors and/or social media users, and the stage on which we will focus today, a content analysis of Facebook/Twitter profiles and YouTube channels of four museums in Belgium and in Luxembourg.

3. STABILIZING UNSTABLE DATA AND PREPARE A CONTENT ANALYSIS OF BIG DATA

We will concentrate our paper on the preparation of our content analysis. The content analysis was the most arduous part of our research although the two other stages (surveys and interviews) weren't smooth sailing either. Most of the time during a doctoral research, a fail is not a real fail but an interesting signal about the core of the materials or the central research question. All the problems that we faced as far as the content analysis is concerned were about the specificities of the Internet, compared to the other media, about museum missions and audience's expectations or silences.

We began our content analysis with an observation of museum institutional websites and their exclusive features. We soon found out there was a hitch: where could we find audiences on a museum institutional website, except in the web analytics who are rarely and not easily shared by the museums themselves?

In 2012, Facebook was on its rising curve and many initiatives were mixing museums, audiences and social media (for instance 'Ask A Curator' on Twitter, 'The Commons' on Flickr, etc.). There was a lot of social pressure on museums, as being on social media was sometimes an opportunity, most of the time an obligation, especially for the small and medium sized museums. So, we chose social media as the field of our doctoral research. We selected the most popular social media of the time for museums and audiences: Facebook, Twitter and YouTube. And let the challenges roll on!

A content analysis begins with the data collection. Or rather, with the reflection on the big data collection process. What did we want as content analysis materials?

From the start of our research, we chose to work with medium sized museums. We worked with four museums, two in Belgium and two in the Grand-Duchy of Luxembourg. In these four museums, there are two (relatively) new museums (2007 and 2011) and two museums have a long history behind them. The four museums are part of the art and history museums family, with an 'intruder' which is a contemporary art museum. All the four museums are important for their geographical area but remain on a human scale. The idea was that the scope of the data would be important, but not as that of a superstar museum (e.g. Musée du Louvre or MoMa).

Our observation period runs from 01/01/2011 to 29/05/2012. The beginning of the observation was determined by the opening of one of the four museums and the arrival of another museum on Twitter. We wanted to analyze at least one year of 'daily publishing', the end was the day the IT researcher who helped us began the data extraction.

If we had used to collect data to achieve media content analysis, the challenges of these big data collection were rather new. One of the most problematic issues with the social media is the instability of the data. At first, we tried to extract it with the Firefox add-ons 'ScrapBook' but the result was incomplete and it was difficult to find where the missing data was. We also thought to build our data archive with a copy/paste of the texts and the screenshots for the images and videos. The problem with a screenshot is that we loose the links and the 'substance' of the Internet [10]. The texts are more exhaustive but the pictures and videos are crucial on Facebook, Twitter and, of course, on YouTube. We worked with an IT researcher who usually builds up the Linguistics corpus. We submitted a list of data which were important to our research… and we received a data (and metadata) archive with more than 5000 files, including 2471 tweets, 152 YouTube videos and 309 Facebook posts. The other files contained the comments, the pictures and the videos' metadata.

The extraction has been conducted in two phases, in May and September 2012, due to challenging technical conditions. We do not know all the secrets of the extraction but we do know that it was laborious. We found the building process difficult but really helpful to our doctoral research. Internet is made from various materials and our data collection process reflects perfectly this characteristic. More than three years later, we are relieved to have spent so much time collecting these big data and to have archived them in an as sustainable way as possible.

If Facebook massively harvests our data, this social media does not generously share them. Despite this strong and meticulous construction of our data archive, the original data has changed, posts have disappeared, some links were dead, and Facebook changed its design. We were able to make up for these issues with screenshots and the data extraction, but some data is still missing, except that this time we know at least where it is.

The NVivo add-ons 'NCapture' has been launched in 2013 and we've tried to collect some data from Twitter profiles with 'NCapture' to compare with our data archive. We found that our data were more complete. For example, on Twitter, the metadata of a retweet gives the primary author as author, which seems completely logical. But if the automatized extraction was only based on the museum as author, precious data for our content analysis were lost: a retweet of a user tweet by a museum is a trace of the participation of a user.

The 'data architecture', the way we've thought the building of our data archive, helped us a lot, as well as the metadata automatically extracted by the IT researcher. Each file has a unique tag (social media + name of the museum + ID numbers given by the social media), which was convenient for the anonymisation of the comments… or searching a specific tweet in a tweetline which is no more easily findable.

After this long process, we had our steady data archive. The next challenge was to find a way to analyse and interpret such a large amount of information. We tried at first to code the data with the NVivo software. We've met several technical issues, firstly due to the pragmatical fact that our department is working on Macintosh devices and that in 2012, NVivo was only available on Windows. We solved the problem but the data were quickly heavy. We had a large data archive and much more of coding information. The output of the coding information was extremely difficult to verify and interpret. It became impossible to have a global vision of it. NVivo is a wonderful tool for many researchers but in our case, we decided to code the big data in an Excel shift and do the quantitative inquiry with SPSS.

Done manually this coding from big data and metadata was difficult but the best way, we thought, to adjust the coding, to know better which information was helpful to our research and which data wasn't so important. Working on an Excel shift allowed us to complete the data: for example, a translation of the content (our four museums speak in five different languages), doing the distinction between the publisher and the primary author of a content, having a quick way to check the consistency of our coding, etc.

We started the content analysis at the same time as the *in situ* and online survey data analysis. The content analysis was an interesting way to fill the gaps left by the other parts of our research. The surveys and the interviews talk about the reported uses of social media by the museums and the audiences, our content analysis is a way to give voice to social media, to temper or confirm the hopes placed in social media.

4. FEW THINGS THAT WE LEARN WITH USING BIG DATA IN A SOCIAL SCIENCE RESEARCH

The data archive was, of course, not entirely perfect and we had to mourn the loss of some information, like the sociological or personal data of the users who wrote comments. Apart from the ethical problems involved, we could not spend more time on our content analysis. The other steps of our research make up for this lack of information but we had to put limits to put an end to our research.

More than in other contexts, finding the right balance between too much data and losing some interesting data is the biggest challenge of big data. An extremely complete data collection is sometimes useful in social science, in order to open new pathways outside the expected methodological choices but we, as researchers, need to be cautious and moderate on working with the amount of data we face.

For example, given the choice again, we would not take such a long duration of observation but several periods of 2 or 3 months, on diversified situations. Working on shorter periods would have decreased the size of our data archive. It would have been possible to do so as our research questions permitted it and because media planning in the museums do not change quickly. This solution is unfortunately not relevant in every situations. As we've seen earlier, we depended on the work of another researcher, a situation which could not allow an adjustment of this methodological choice.

If participation is an important concept in our research, the experience of our research in data collection and analysis of big data showed us that it has become essential to work together. Big data requires a joint effort, in order to know the skills of researchers in other fields and to combine different expertise. The needs of an IT researcher or a Social Science researcher are different but in our respective research we are facing the same challenges regarding big data.

5. ACKNOWLEDGMENTS

Our thanks to Hubert Naets, IT researcher at the Université catholique de Louvain, who helped us by collecting the social media data and to Agnès Calatayud who proofread this paper.

6. REFERENCES

[1] Davallon, J., *in* Schall, C., Colas-Blaise, M., Tore, G. M. 2014. *Parlons musées ! Panorama des théories et des pratiques*. Editions Guy Binsfeld, Luxembourg.

[2] Macdonald, S. 2006. *A Companion to Museum Studies*. Blackwell Pub., Malden.

[3] Parry, R. 2010. *Museums in a digital age*. Routledge, New York.

[4] Bilandzic, H., Patriarche, G. 2012. *The social use of media*. Intellect Ltd., Bristol.

[5] Curran, J., Fenton, N., Freedman, D. 2012. *Misunderstanding the Internet*. Routledge, New York.

[6] Jenkins, H. 2008. *Convergence Culture: Where Old and New Media Collide*, New York University Press, New York.

[7] Millerand, F., Proulx, S., Rueff, J. 2010. *Web social : Mutation de la communication*, Presses de l'Université du Québec, Québec.

[8] Carpentier, N. 2011. *Media and Participation. A site of ideological-democratic struggle*, Intellect Ltd., Bristol.

[9] Carpentier, N. 2014. *Engagement social, société civile et médias en ligne*, Fondation Roi Baudouin, Bruxelles.

[10] Brügger, N., *in* Consalvo, M., Ess, C. 2011. *The handbook of Internet Studies*, Wiley-Blackwell, Malden/Oxford, 27-36.