Loading… The Journal of the Canadian Game Studies Association
Vol 10(16): 87-108
http://loading.gamestudies.ca

Using Big Data Tools and Techniques to Study a Gamer Community: Technical, Epistemological, and Ethical Problems

Maude Bonenfant
Université du Québec à Montréal (UQAM)
bonenfant.maude@uqam.ca

Fabien Richert
Université du Québec à Montréal (UQAM)
richert.fabien@uqam.ca

Patrick Deslauriers
Université du Québec à Montréal (UQAM)
deslauriers.patrick@courrier.uqam.ca

Abstract

This paper discusses an exploratory approach taken by researchers in the fields of semiotics and communications in order to not only share a specific research experience, but also help build a research sector that combines game analytics with social sciences. The main objective of our research was to define parameters of digital identity within the framework of the study of an online video game player community. To this end, we examined several constitutive elements of digital identity, namely the effects of the "avatar" apparatus on the identity of users, online interactions, and the meaning of "living together" in the digital world. We used both qualitative and quantitative methodologies: a semiotic analysis of the game, a discursive analysis of the forum, semi-structured interviews, and an automated analysis of big data sets. In this paper we will focus on the automated analysis of big data sets, addressing two key points: the working method developed by the research team, and the achievement of the research objectives by merging quantitative and qualitative perspectives together. Following a summary of the research approach, this article will present the methodological, epistemological, and ethical difficulties that may be encountered in studying a player community with this type of research approach.

Author Keywords

Big Data; Game Analytics; MMO; Data Mining

Introduction

The production of digital data is quickly and vastly expanding, to the point that this phenomenon can now be called "big data." Big data can be characterized as the production of massive amounts of data and the development of tools and techniques that facilitate data stocking, processing and analysis. Big data's most innovative features are the cross-referencing of different types of databases (structured, semi-structured, and non-structured) and the use of advanced correlation methods in order to reveal "patterns" that would otherwise remain

invisible (Bollier, 2010). The videogame industry is not immune to this phenomenon: an increasing number of companies have started gathering a massive amount of information by tracking player behaviour in order to improve gaming experiences and increase profitability. The videogame industry currently recognizes the importance and the competitive advantage of using data. This creates a need for data analysts and data scientists to extract meaning from the data – that is, to translate computerized data into intelligible and useful information that will help companies meet their objectives. Data may initially be collected and analyzed using a range of automated tools, but human intervention is still required for in-depth and precise analysis.

Big Data as a method of knowledge production is gaining popularity not only in the videogame industry, but also in the academic world and in game studies more specifically. However, while solid expertise in this type of methodological approach based on big data and high performance computation has developed in certain fields (genomics, for instance), social sciences are not traditionally specialized in handling both this amount of data and the computer tools required for their analysis. Although some researchers have stood out in this type of research (Ducheneaut, Yee, Nickell & Moore, 2006; Williams *et al*., 2006; Ducheneaut, Yee, Nickell, & Moore, 2007; Williams, Yee & Caplan, 2008; Drachen, Canossa & Yannakakis, 2009; Lewis & Wardrip-Fruin, 2010; El-Nasr, Drachen & Canossa, 2013, etc.), expertise in this expanding field of study still requires elaboration. Building a solid and reliable methodology requires sharing working methods.

Several scholars in the field of game studies have already contributed to such developments (Bainbridge, 2007; Consalvo & Ess, 2011; Boellstorff, Nardi, Pearce & Taylor, 2012, etc.) and this research stems from and builds on those advances. The necessity for an epistemological and ethical, if not ontological, framework for thought on this topic is all the more evident considering that:

> in digital games research we are (…) looking at an unprecedentedly extensive and intensive data source. We are, as well, examining a phenomenon, mass-scale playful inhabitation of virtual worlds that has never before taken place. In other words, both substantively and evidentially, our field is without precedent.
>
> de Castell, Jensen, Taylor & Weiler (2012)

Therefore, rather than showing the results of the automated analysis of data produced by a particular online videogame player community, this article reports on the exploratory approach taken by researchers in the fields of semiotics and communications in order to not only share a specific research experience, but also help build a research sector that combines game analytics with social sciences. Consequently, this article will discuss two points: first, the working method developed by the research team and the achievement of the research objectives by merging quantitative and qualitative perspectives together. Following this summary of the research approach, this article will present the methodological, epistemological, and ethical difficulties that may be encountered in studying a player community with this type of research approach.

Research context

Big data processing and analysis techniques in the context of videogames derive from the fields of statistics and computer science. Applied to game analytics, this expertise is usually intended to assist decision-making, improve gameplay mechanisms, or model players' behaviour, based on the traces they leave through their in-game activity (El-Nasr, Drachen & Canossa, 2013). Game designer David E. Kennerly (2003) has identified the advantages of data-mining techniques, especially in massive multiplayer online games (MMO). Moreover, certain types of casual games, especially "pay to play" and "social games", are mostly data-driven, as data processing allows improving some game features and functionalities. According to McCallum and Mackie (2013):

> The rise of social gaming has created a new model for game development and publishing. Rather than developing a complete product and releasing a full game, social games are usually released early as small games that are then incrementally built as the game becomes popular. This incremental improvement to the game can be informed by players' actions in the current version of the game. This data-driven approach (Rabin 2000) emphasizes the use of data to understand and improve the game, rather than merely use the intuition of the designer (Isbister and Schaffer 2008; Pagulayan et al. 2003; Derosa 2007).
>
> *in* El-Nasr, Drachen & Canossa, p. 178 (2013)

Within this context, data analysis directly impacts future game development, as results appear to give direct answers to game developers (p. 6). The main argument brought forward by this approach is based on the idea that gaining direct access to player information increases the veracity of results (Kennerly, 2003), and has often been used by other researchers. Yet, one of the main problems data scientists are faced with is the way meaning is produced through data. Thus, much research has focused on the development and improvement of visualization techniques that both enable the analysis of videogame design problems and provide a better understanding of players' behaviour (Hoobler, Humphreys & Agrawala, 2004; Magerko & Medler, 2011; Moura, Seif El-Nasr & Shaw, 2011). Visualization techniques are useful in reporting data on players' movement (using heat maps, for instance) and tracking different in-game events or actions. In order to shape player behaviour, other techniques that are specific to statistics, such as neural network and machine learning, are prioritized, which requires expertise in computer science (Charles & Black, 2004; Drachen *et al.*, 2012).

While some research primarily focuses on improving gaming experiences (by collecting data) or enhancing game mechanics (for example, by adjusting difficulty levels in accordance with the overall performance of players), the bulk of recent research aims to shape player behaviour. Selecting MMOs as case studies makes it easier to achieve this goal, and these games offer great potential for deep and complex study of interactions between players. For instance, *World of Warcraft* (WoW) has been the subject matter of many behavioural studies interested in gaming communities. For example, between 2004 and 2009, Christian Thureau and Christian Bauckhage (2010) conducted a major study of player behaviour in guilds (or teams). The authors collected connection logs from players and guilds (192 million recordings for 18 million players and 1.4 million guilds) with the aim of categorizing the different types of guilds and visualizing their progress. Similarly, other researchers have examined different elements that determine a guild's success (Poor, 2015; Ducheneaut *et al.*, 2007), how classes, objects, and equipment are selected (Lewis & Wardrip-Fruin, 2010), and player performance through the analysis of game time, learning curves, and reasons for quitting (Ducheneaut *et al.*, 2006).

Computer programs developed specifically for WoW assisted all of these studies, as they allowed collecting information about players while they played on a game server.

Researchers have also focused on *Star Wars Galaxies* (Ducheneaut & Moore, 2004), *Everquest II* (Williams *et al.*, 2008 & 2009), and *Tomb Raider Underworld* (Drachen *et al.*, 2009). In the same way as for studies of WoW, research on *Star Wars Galaxies* required the development of a specialized computer tool, which was used to collect data on a variety of in-game interactions between players (by using the "/log" command). This data collection method was rather challenging for researchers since they had to create avatars and stay logged in for long periods of time. As for research focusing on *Tomb Raider Underworld*, which studied the visualization of player behaviour based on in-game performance, live data was tracked using an *Eidos* computer software program (*EIDOS Metrics*). In the case of *Everquest II*, Sony Online Entertainment provided access to the game's database, and researchers focused specifically on player profiles as well as gender stereotypes. In short, all of these studies were conducted in collaboration with video game companies, which enabled and facilitated research.

Overview of methodological approach and research conducted

Our own research situates itself within that context. It was conducted between 2013 and 2015, with the support of a Social Sciences and Humanities Research Council (SSHRC) Insight Development federal grant. The objective of this SSHRC grant program is to promote "the development of new research questions, as well as experimentation with new methods, theoretical approaches and/or ideas"[1]. Therefore, from the onset in 2012, our research has been exploratory in nature. The fact that this field of research was still relatively new, as shown by the literature review presented above, confirmed the need for an experimental approach.

The main objective of our research was to define parameters of digital identity within the framework of the study of an online video game player community. To this end, we examined several constitutive elements of digital identity, namely the effects of the "avatar" apparatus on the identity of users, online interactions, and the meaning of "living together" in the digital world. We used both qualitative and quantitative methodologies: a semiotic analysis of the game, a discursive analysis of the game's forum, semi-structured interviews, and an automated analysis of big data sets. In this paper we will focus on the automated analysis of big data sets. The goal of this big data-based methodology was to identify and map a community's distinctive set of behaviours with the help of indicators focusing on in-game identities and actions.

Our research examined a game produced by *Vandal Games*, a Montreal based start-up, which permitted us to work with a copy of their database. This database mainly consists of four types of data: player characteristics (gender, home country, gender of the selected avatar (permanent)), player actions (number of games won or lost, in-game money earned, class of the selected avatar (temporary)), and information on game sessions (match date and time, match duration, number of players involved). Other types of information also formed part of the data tracked by *Vandal Games*: *Google* analytics (connection data) and *Facebook* analytics (data aggregation and player identification, since the game under study is played on *Facebook*).

The game that we decided to study - *Big Story Little Heroes* (2012) - is one of the first real-time videogames available on *Facebook*. Its rules are reminiscent of "capture the flag" games and partly reflect the organizational logic of MMO games, in which one controls an avatar and plays in a pre-determined world. Each match involves two teams of six players. A team's objective is to steal the "statue" (which acts as a flag) from the enemy's base. Before a match can start, each player has to choose a class (archer, mage, priest, engineer, warrior), which

determines the abilities, strengths, and weaknesses of the player's avatar. Once the game has started, players can win the match by fighting their enemies, both human and non-playable characters that are controlled by the game's computer code, and by completing the main objective of stealing the "statue". When the match is over, a set of statistics sums up each player's performance: the number of kills (enemy players), number of deaths (one's own avatar), amount of money and experience points (XP) earned, etc. Matches last ten minutes on average. Between each match, players may access a menu in which they can customize their avatar, talk with other players, and buy items (weapons, armour) in the game's shop. In short, the game replicates classic features of MMOs such as gameplay and possible actions in the menus, but in-game mechanics are taken from "capture the flag" games.

Once the agreement with *Vandal Games* was finalized and access to the database was granted and shared, the research team had to familiarize itself with the game, the database, and the computer tools. Since the company set up the database, the research team had to learn about both the attributes and the data that had already been collected. The 34 tables that are directly linked to *Big Story Little Heroes* players are in SQL format and coincides with a specific timeframe (from June 10th, 2013 to May 22nd, 2014). For the purpose of this research, we used *EasyPHP*, which allows installing an *Apache* server on a computer. We needed two other interfaces or platforms: a *MySQL* database management system and a dedicated interface (*PhpMyAdmin*) for the easier administration and management of recordings within the database. The latter Web development platform is appealing because it offers the possibility of working locally without having to connect to an external server.

After installing the database in a local server, the data was then explored with the help of a software program (*RapidMiner*) adapted to big data searches. We mainly used this software to clean the data, send requests, and correlate specific tables with each other. *RapidMiner* is written in *Java* programming language, and offers a graphic user interface (GUI) that allows completing different tasks and understanding a variety of operators that are specific to machine learning, data mining, text mining, and predictive analysis techniques. Although visualization through *RapidMiner* was tested, the results were mostly presented in charts with *Excel* tables.

Cross-referencing data types and results

It should be stressed that the quantitative results were cross-referenced with qualitative results. More specifically, most of *Big Story Little Heroes*' official forum was analyzed through content and sentiment analysis. The aim of this analysis was to understand how players felt about the game, what their own experience of it was, what changes *Vandal Games* implemented, etc. This way of working with mixed methodologies helped identify significant differences or irregularities in quantitative data. It also helps find complementary information and explanations through qualitative data analysis. Indeed, identifying significant elements in qualitative data then allows verifying these elements in quantitative data. In other words, information collected outside of the quantitative database helps detect singular markers or particular events. Afterwards, these findings can be correlated with the results obtained from the analysis of the quantitative data.

For example, we identified the most active players on the official game forum (for the whole period of time studied) with a manual, precise and thorough encoding process of this web platform. These players were then identified in the game; their behaviour was analyzed and specific behavioural traits were categorized in accordance with our criteria and tags. The idea was to consider these players as community leaders, to classify their interventions on the forum (sarcasm, humour, boastfulness, opinion, question, etc.), and then analyze their in-game

behaviour. Conversely, we wanted to identify the most active players in the game and analyze their behaviour on the forum.

Main difficulties encountered by the research team

The first challenge we faced in this type of research was the attempt to tie the interests of the industry and those of the academic field together. Despite the fact that daily interpersonal interactions with *Vandal Games* employees were cordial, if not friendly, the company's objectives, working methods, and relationship to time and knowledge were quite different from those of the research team. This required that both parties adapt to each other. In this regard, the challenge of good planning is to better invest communication channels so that everyone can reach their objectives in a trustworthy environment. Consequently, trust must also be established from the outset and must stem from a mutual understanding of individual and collective goals. Researchers ought to keep the stakes of industrial secrets in mind so that their work does not negatively impact the company, but they must also maintain their own research objectives and their academic freedom. Indeed, one should not dismiss the risk of altering the study, consciously or unconsciously, so as to meet the company's needs. While we agreed to give advice on different issues including game design, elaborating knowledge and enriching our field of study have remained primordial ethical principles in the research team. Meanwhile, on the corporate side, our scientific objectives had to be kept in mind as our main priority.

Such efforts made in order to not lose sight of scientific interests were especially important considering that the research team always remained dependent on the company and the data it agreed to share. For example, we were denied access to monetization data. While it cannot be simulated in lab, researchers can only access this type of data if they work with a company that agrees to either share a current database or provide a copy of an old one. They can also try to access open-source databases, if it is possible. Furthermore, since the company builds its database and selects what to track, the research team relied solely on the available data. This will be the case for any research team unless the company one works with agrees to include trackers in the game's code itself. The fact that the company produced our main research material was another challenge for the research team: understanding what particular data refer to and figuring out how to make sense of the data was challenging. If sufficient information about the data's attributes and organization are not produced or provided, it can become quite complicated to identify what the data represents. Such lack of information on the meaning of the data highlights the necessity for good communication between the company and the research team. This is especially relevant if the company starts tracking new attributes as the game and its database evolve.

The data cleaning process is an even more crucial, yet tedious and time-consuming task. 'Cleaning' means that before analyzing a database, researchers have to eliminate the erroneous data that could skew the results. For instance, in our case, some players seemed to have played their last match (last login) before their avatar was created (creation time); this situation is undoubtedly impossible since the creation of an avatar necessarily predates the last match played. Thus, this type of unusual and erroneous data has been removed from the database. Similarly, during a research led by Chris Lewis and Noah Wardrip-Fruin (2010) on more than 100 000 avatars in WoW, such errors were identified. For instance, some data had a negative value: it is impossible for a player to score -5 kills in a match, as that value should be equal or superior to 0. In this situation, as the authors wanted a valid sample, they eliminated problematic data from the database. However, Lewis and Wardrip-Fruin kept some 'abnormal' data in their database because they had no 'proof' of their incorrectness: for instance, a player

reaching level 80 in a short period of time. During our research, we were also confronted with abnormal and uncertain data. For example, a player gathered a lot of in-game money while playing relatively few matches. Nevertheless, as Lewis and Wardrip-Fruin did, we kept such unusual data in our sample since we could not prove or disprove its veracity.

Adding another type of data - the in-game data produced by the *Vandal Games* employees – also could have distorted our results to a certain extent. It is important to specify that this type of data has not significantly altered our results since the number of employees playing the game remains small compared to the size of the community under study. Yet, we had to consider that those employees were playing with obvious advantages, as they had very accurate knowledge of the game's stats and mechanics, access to every item and class, as well as the possibility to change attributes/functions during a match, among other privileges. As these prerogatives enhanced employees' performances, the data they produced was misleading: it did not represent a 'normal' match between equal players. However, considering the sheer amount of data under study, it was rather difficult to differentiate between employees and 'normal' players. Therefore, even though a certain number of these employees' data was removed from the database (based on the list of pseudonyms the company provided us with), we could not rule out the possibility that this type of misleading data might have influenced our results, particularly upon evaluating players' performances.

In the same way, bots could have skewed our results (whether we have proper evidence for this or not). Bots (contraction of the word "robots") are computer programs that communicate with servers in order to perform actions. Moreover, players can use of a virtual private network (VPN) to change their geographical location and they are free to select a gender that is not their own when signing up to *Facebook*. For that reason, the question about the validity of data seems fundamental: is this data 'valid' and does it meet the conditions of its functionality? If we agree that this data truly exists, what does it represent from the point of view of a computer system, versus what is really sent? It is impossible to interpret the results as long as the data under study has not been checked and verified in order to understand its object of reference.

Aside from the validity of data, the format in which the data is presented is another problem. For example, the data from the quantitative and qualitative databases cannot be cross-referenced right away since the nature of the data itself and the software and extensions used (file types) are not equivalent. For instance, while we studied and measured individual performances (quantitative approach), we described and characterized types of interactions between players (qualitative approach). Thus, our two databases did not initially match. In order to remedy that, we established objectives that would facilitate the cross-referencing of these databases. As our goal was to develop a general picture of the most active players on the forum, we looked at the forum's database, which we had analyzed using a qualitative approach, to identify the players that had posted the most. The forum is independent from the game, and pseudonyms on the forum did not always match pseudonyms in the game, which meant that successfully linking players as they appear in-game with their respective names on the forum was challenging. Conversely, we could not verify whether a player using the same username in-game and on the forum were actually the same person.

Consequently, cross-referencing databases can be an appealing, although near impossible task. The validity of cross-referencing will never be assured for two main reasons: 1) the absence of a relationship between data collected in-game and data collected on out-game platforms; 2) the fact that these two spaces are not managed by the company. By way of illustration, aliases in *WoW* are directly linked to the aliases used in the game's official forum; this means that cross-

referencing in-game and out-game behaviour is easily done. So, in the case of our own research, to be able to cross-reference the results from two independent but linked platforms with certitude would have helped us better understand the community. Indeed, a number of authors reiterate the necessity of cross-referencing quantitative and qualitative databases for many reasons such as understanding motivations (the why) behind behaviours (the what or how), as well as identifying players who play differently (Ducheneaut *et al*., 2006; Tychsen, 2008; Lewis & Wardrip-Fruin, 2010).

An additional challenge raised by some authors relates to the researcher's familiarity with the game studied (Wood, Griffiths & Eatough, 2004; Ducheneaut *et al*., 2006; Williams, 2010). In order for research to run smoothly, it is often necessary for the researcher to meet this criterion. Knowledge of the game should be acquired prior to collecting and analyzing data. In other words, researchers have to dedicate several hours of their time to simply playing the game before they are able to start the research. In doing so, researchers can better understand the game's mechanisms and subtleties, gameplay, the several attributes and statistics shown, and the usual sequence of events in a match. They can also get acquainted with the type of community playing. This helps researchers contextualize their object of study, ask pertinent questions, and improve their analysis with a more intimate understanding of the meaning of the results. Indeed, good initial knowledge of the game positively impacts the interpretation of data and the production of results since researchers may remember and reflect on their own experience with the game throughout the analytic process. Dmitri Williams' (2010) perspective on this matter is straightforward: "to understand the impacts that code and culture can have, I insist that any working team have first-hand experience within the virtual world. My rule is simple: If you haven't played it, you can't study it" (p. 13).

In light of this recommendation, each member of our research team created an avatar in *Big Story Little Heroes* and played matches. Despite limited availability, each team-member played the game, though no one reached level 30. Admittedly, studying the game's 'bible' would have been sufficient to understand the main functionalities and objectives, however this would not have allowed us to understand other aspects of the game such as the tactics and strategies used to win, the strengths and weaknesses of each class and item, and the relationships between teammates (communication, mutual aid, individualism, etc.). Our analysis and interpretation of the results were affected by our relative unfamiliarity with the game. For example, a player reached level 30 with an almost perfect win rate (97.5%); another player earned an abnormally important amount of money for each match played. If we can only access directly available information, how should we interpret these two cases in terms of performance and effort? Are these cases of erroneous data, outstanding players, or cheaters? These questions remain unanswered, and a better expertise on the game would have improved our interpretation of the results.

This expertise is especially relevant considering that the definition of the criteria or boundaries of the research sample is an arbitrary choice. For instance, while studying the performance of the 'best' players, we had to set specific parameters to delineate this category ('best'). In other words, we had to answer the following question: what criteria are used to determine whether a player belongs to the 'best' category or not? With regards to our research questions, we had to specify the parameters according to which our samples would be circumscribed. Coming back to the "best" category, we decided that players should have participated in at least 100 matches with a win rate of over 80%. These two criteria emerged from a methodological choice made in line with our own observations of the data and our experience of the game. The decision to include this or that criterion in selecting our sample reflects the arbitrary aspect of the research.

This choice could have been different had other researchers worked on the same data. While it is necessary to set parameters in order to analyze a coherent database, this choice ultimately circumscribes the results within a specific and limited framework.

Those parameters and limitations were also introduced as a way to reduce the amount of data to be analyzed. While we only accessed data pertaining to a rather narrow timeframe, the amount of data in question was still enormous. We also had to work with a database showing different attributes. That massive amount of data ended up being an additional inconvenience in the formulation of research questions. According to many researchers who were faced with similar problems, one has to learn how to format/simplify/clean databases and identify the metrics to be used (Kim *et al*., 2008; Williams, 2010; Hopson, Hullett, Nagappan & Schuh, 2011; El-Nasr, Drachen & Canossa, 2013). Considering the vast array of research questions that can be asked and attributes that can be studied, the need to restrict the size of databases in order to generate significant results is obvious. In our research, regardless of the questions asked, we had to set parameters in order to select the data to be analyzed. For instance, when evaluating player performance, we chose one criterion to delineate our sample, which reduced the quantity of data under study: players had to have played at least ten matches. As seen in the above paragraph, the results that were generated could have gone in a different direction had the arbitrary criteria been different (for example, we could have selected players who have played five or twenty-five matches).

Lewis and Wardrip-Fruin (2010) point to the timeframe selected for analyzing the data as an important parameter. Researchers should select dates in between two major changes made to the game (updates, patches, etc.). This choice of a timeframe would show how changes affect the game and its players. It would also give a good idea of the community's main traits at a specific point in time. For our research, we had access to a database that covered a timeframe during which major changes had been brought to *Big Story Little Heroes*. Not only was the amount of data we accessed massive, but this data also represented the sum of many player experiences spanning several key phases in the development of the game. Consequently, the important quantity of data and the ever-changing context of the game (updates) made it difficult to really grasp the identity of the community.

Some of these problems could have been avoided had we accessed the database earlier. In other words, establishing a partnership with the video game company prior to the game's release would have helped us understand the system, master the game's subtleties, and adequately determine the research's objectives and questions. Moreover, this would have allowed researchers (in partnership with the game developers) to design the database (Williams, 2010). Without this early access, we had to formulate our research questions after data had already been produced. This circumstance added to the other challenges we faced during our research. If we had worked with the company from the outset, before the game's release, trackers could have been added to the tools, for example, in order to help answer specific research questions formulated beforehand.

Finally, the division and specialization of tasks within our research team was problematic in several ways. Firstly, only one researcher worked directly on the database while generating statistical results through mathematical computation. This person thus had to acquire different techniques and types of knowledge (computational, mathematical, statistical, and analytical) in order to fully achieve the goals associated with interrogating the database. Furthermore, from the outset, all members on the research team lacked technical expertise. In a study on massive data on the game *Everquest II*, Dmitri Williams (2010) states the importance of familiarizing

oneself with the software used in managing and analyzing data. Williams actually confirms the need to develop an expertise on these systems before even touching the data. We did not have access to any help from such resourceful professionals.

Additionally, the analytical tasks required the use of a variety of computer software in order to clean, organize, calculate, and visualize the database. The multiplicity of interfaces and analytical steps lengthened the research process and increased the difficulty of learning about the systems' inherent logic and how to use them. Eventually, while we decided that each team-member would specialize on a different part of the research, team-members were unable to fully grasp each other's work and coordinating the team became difficult. For this reason, and since the required technical knowledge was diversified and advanced, methodological descriptions were hard to communicate between members of the research team.

Our experience in learning from this difficulty reveals an essential aspect of this type of research, which is the need to work with computer researchers from an interdisciplinary perspective. Indeed, the necessity to collaborate with experts who can understand and manipulate computer tools becomes obvious since that expertise requires many years of specialization in Computer Science. Our partial technical knowledge limited us in the use of the tools, the analysis of the data, and the coordination of the research team. Nonetheless, this exercise helped us acquire basic skills that have made us better prepared for future research on similar topics. Also, we are now able to collaborate more efficiently with computer science researchers, as we understand the basic logic and vocabulary of big data research. In this regard, Burgess and Bruns (2015) stress that research based on computer data are more than ever "blurring disciplinary boundaries between humanities, social sciences and computer science" (p. 95).

Epistemological problems

Besides the technical problems related to the manipulation of a massive amount of data, this type of methodology raises several epistemological issues. Busch (2014) identifies twelve challenges to keep in mind while working with large-scale data sets. Boyd & Crawford (2012) highlight four epistemological caveats to consider:

- Big data changes the definition of knowledge
- Claims to objectivity and accuracy are misleading
- Bigger data are not always better data
- Taken out of context, big data loses its meaning

Summing up these precautions, we may ask the question: with these computer tools, these analytic principles, and this amount of data, what kind of knowledge do we produce, and what are the limitations of such knowledge? While Chris Anderson provocatively said in 2008 that "the data deluge makes the scientific method obsolete" and that this means "the end of theory", it seems that we should ask not whether theory is useless in this era of big data, but rather which theories frame and emerge from research using big data methods. Put differently, which presuppositions can we identify in order to really understand this mode of knowledge production? In line with Crawford, Milner & Gray (2014), "we argue that big data *is* theory".

Indeed, Anderson (2008) proclaims the end of theory because he considers that scientific methodologies organized around hypotheses, models, and experimentation are outdated: "We can stop looking for models. We can analyze the data without hypotheses about what it might

show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot." According to Anderson, numbers speak for themselves and meaning emanates directly from data. Data would thus answer questions that we don't even have to formulate anymore (or rather, that we will only ask after the fact). It looks as if the world could be explained strictly in quantified patterns. More specifically, in the case of our study, it is as if data produced by players' activity could explain players' behaviour, social relationships, identity dynamics, etc.

If the world could unfold in real time and 'speak of itself' before our eyes, there would be no need for theories anymore. This would mean that data would be in direct contact with 'reality', as if there were no mediation between researchers and their object of study. This represents the paroxysm of a system of thought based on 'facts' in which only objects and actions (neither true nor false in themselves) constitute the whole spectrum of knowledge. In other words, within this framework, gathering all the 'facts' (presented as multiple data) becomes the foundation of knowledge itself. Seeing as theories have become useless at best, and deceptive at worst, it seems that one could only reach truth through facts – through data. To illustrate this line of argument, we may refer Greek mathematician Euclid's *Data*. The term "data" (*δεδομένα* in Ancient Greek) is the book's single-word title. In this work, Euclid aims to explain how information that is *given* (literally translated as *data* in Latin) is to be understood within the framework of geometry, which is the science of 'what is known' and 'what is attested as such' with regards to the measurement (*μέτρον*, -metron) of the earth (*γῆ*, gê-).

Nevertheless, this way of looking at big data dismisses the idea that all human formalization implies mediation. Data is not reality itself; it is a technological representation/mediation of reality that relies on the translation of knowledge and actions, transmitted through electronic equipment, into computer language. As such, historically framed power relationships remain, even at the level of simple information (Foucault, 1971 & 1976). Many authors (Ellul, 1954; Simondon, 1958; Mumford, 1967 & 1970; Stiegler, 1994, 1996 & 2001) have demonstrated the power that technology has on the future of humans and society. In that sense, technology's apparent neutrality is a lure: ideologies, values, beliefs, and representations of the world are inscribed within technological apparatuses. These unstated presuppositions affect our understanding of the world and our behaviour. Thus, they shape our knowledge about player practices and the *Big Story Little Heroes* community. For example, it is important to remember that the database studied was built entirely by *Vandal Games*, and serves as a way to further the company's best interests. The formal characteristics of data, the way in which it is collected and treated, and the values that intervene in their processing can all be questioned (among other things). To recall the title of a book written by Gitelman (2013), "raw data is an oxymoron" and the metaphor of data mining is misleading. Indeed, data is not a commodified natural resource "to be controlled", nor is it an object of consumption or "nourishment to be consumed" (Pushmann & Burgess, 2014). Rather, data is an interpretation and "a limited representation of the world" (Loudon & al., 2013 *in* Boellstorff, 2013).

Yet, the truth-value of the results that big data analysis tools generate is significantly higher than that of other representational and organizational modes of knowledge production. This truth-value is even greater in the case of massive data since the logic of big data implies that the more data is collected, the more genuine and accurate results should be. It is as if quality became an emerging feature of quantity (Cassin, 2007). Following that logic, the ultimate fantasy would be to translate the whole world into digital data: more data would surely mean stronger patterns, easily identifiable irregularities, and better (more precise) predictions. In other words, results acquire greater truth-value as the quantity of collected data increases, until

it is ‘too big to know’ and we start believing that our intelligence is proportional only to the size and extent of our networks (Weinberger, 2014).

We imagine data to be neither good nor bad, but simply an accurate reflection of reality because data is said to be free from ‘biased’ and ‘fallacious’ human subjectivity. But this belief must be understood in the context of its Ancient Greek roots rather than its most recent iteration with the arrival of computer systems. Assuredly, in order to fully understand our relationship to big data, we have to refer back to Plato’s warning that our senses are misleading: only with ideas can we get out of the cave and reach the truth. This quest for the truth spans the history of thought, especially in the Western world, in which reason and logic oversee the modes of structuring ‘true’ arguments. For many centuries, God was the standard and ideal through which truth could be understood. From the Renaissance period and continuing into the Enlightenment, divine authority was progressively replaced by science in determining what was true or false. The truth-value associated with science thus increased and mathematics, the ‘queen of sciences’, imposed itself as a model for reasoning. The dream would be for humans to be able to say everything in codified terms (zeros and ones); this way of speaking would be articulated in formal logic and through resolutely true and indisputable mathematical formulas. For instance, thinkers like Gottfried Wilhelm Leibniz wanted to find the Adamic language. Ultimately, the idea was to tackle philosophical problems with mathematical demonstrations: “let’s calculate to settle disputes: this is what arbitrary marks, which we have used until now, wouldn’t be able to do” (1698, p. 151).

Nowadays, the use of mathematics in computer systems (especially with algorithms) perfectly illustrates the idea of surpassing individual and cultural particularities with the help of supposedly true, and thus univocal, results produced by massive data and big data tools. In this context, since mediation has seemingly subsided and technology has come to be considered as a neutral tool, we appear to be exempt of human subjectivity when it comes to producing knowledge. In particular, it seems possible to reduce or even eliminate polysemy. In this way, it looks like we can surpass the limits of the brain in terms of reasoning, data processing, and memorization.

This is what many researchers put forward, while arguing that computer tools and game metrics increase objectivity and truth-value. In this regard, Tychen (2008) says:

> Metrics provide objective data on the interaction between players and games […] Metrics are objective; can be collected in large numbers and map to specific points in a game. In comparison, player-based feedback has much less resolution and is inherently biased due to individual preferences
>
> P. 2

Similarly, Kim & al. (2008) argue that “[f]inally, by standardizing the method, predetermining the analysis, and eliminating interaction with the researcher, the method would have better reliability”. Some scholars seem to affirm that, in that sense, our knowledge about the world would get close to perfect: “[t]he companies that run these virtual worlds typically store all or some time range of the actions carried out within the space. If these could be accessed, they would be nearly perfect unobtrusive data” (Williams, 2010, p. 3); hence, the belief that qualitative data, or any kind of data collected and processed by human researchers, are less valid than results produced via computational tools and data.

Nonetheless, if computer tools for massive data processing provide information about truths (in the plural), data only has value relative to what it represents, in a circumscribed way and

bearing in mind the possible fallouts we have pointed to in the previous section. It is crucial to understand that the data sets that we accessed for the purpose of our research on *Big Story Little Heroes* are only the representation of certain types of information selected by the company. The selection of what is to be considered as data is significant in itself. Indeed, in our case, that process was motivated by the company's own goals, which do not necessarily align with scientific objectives. Arbitrary, subjective, and sometimes faulty choices could have been made, which would determine what information ended up being sought after. The company's choices also neglected certain types of information that could have been tracked, such as data regarding friends lists (identity of players on the list, date on which they were added) or chat room data (number of characters sent, date, channel used). More importantly, data selection leaves an important amount of information about gaming practices behind, as it dismisses untraceable or unquantifiable information. For instance, the fact that the company chose to collect mainly gameplay-related data rather than information on communication between players revealed that players' actions were considered to be more valuable than socialisation processes – which was our main focus. Additionally, the meaning of some data might change depending on its context of interpretation: data undergoes recontextualization between the data collection phase, for the purpose of which some significant elements were selected, and the processing phase, during which the same data is linked with other data regardless of the context of data collection. This data is thus reinvested with a new value that is specific to the mining phase.

In this sense, the truth-value of particular data is relative to the context of what it is supposed to represent or what is possible to represent. According to Christine Borgman (2015), the first characteristic of data is that it represents an observation or an object, that is to say a concrete entity, such as text on a piece of paper or signals sent by a sensor. For Borgman, entities only become data once "someone uses them as evidence of a phenomenon, and the same entities can be evidence of multiple phenomena" (p. 28). In other words, data represents specific information selected by individuals and tools, within the scope of the possibilities they offer: some information is not transposable into data because we do not have the proper sensors yet or because they are not translatable into computer language. Despite the fantasy of big data's totality, that is to say the fantasy of its ability to gather the entirety of possible data and thus represent the whole world at once, not everything can be translated into quantifiable, malleable, and operationalizable data:

> Data is not a mirror of the social; it implies the abstraction of everything from thoughts, emotions, and facts into sets of computable symbols. What is being compromised through such translation? What is lost when the richness and complexity of the social is abstracted into data?
>
> Langlois, Redden & Elmer, p. 7 (2015)

In line with philosopher Jacques Rancière's thought (2003), Galloway (2011) asks: "[a]re some things unrepresentable?" If it can be argued that technologies are constantly improving and that the world is increasingly being digitized into data, the sum of each and every discrete data will never equate a continuous whole. As Anderson (2008) pointed out, considering knowledge as the gathering of data can have widespread epistemological implications informed by computer processing, and affect our outlook on video game players' practices and communities.

Indeed, big data technologies promote, allow, or prevent specific data processing methods. They can even normalize research strategies or naturalize a certain idea about what constitutes knowledge. For example, they already shape the way player behaviour is represented

(necessarily with an 'entry'): each attribute must have complete, normalized, and standardized entries in accordance with a limited number of variables. With better regularity in the entries and variables used, the validity of the data increases. This way of organizing information through predetermined and fixed categories already reveals a particular manner of representing knowledge that is not necessarily the same as the one we might see in social sciences. Computational data evidently gives us hints about player behaviour, but, as seen above, it cannot provide or make up final knowledge about experiences, social relationships, identity dynamics, etc. This is especially the case since Lazer, Kennedy, King & Vespignani (2014) note that "[t]he core challenge is that most big data that have received popular attention are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis". The authors also remind us that the platforms that collect the majority of data belong to corporations. Andrejevic (2014) also highlights the fact that the relationship between the people who collect the data and the people whose data is being collected is asymetrical, which implies a disconnect between the two groups. Another dichotomy exists between those who may access the databases and those who may not: "limited access to Big Data creates new digital divides" (Boyd & Crawford 2012). Indeed, these databases mostly remain companies' property, and so researchers cannot share them in order to promote diversified arguments and knowledge production based on the same sources. Furthermore, results are always-already informed by different elements: what the tool detects and selects, the way in which that detection and selection takes place, and the way information is codified and processed. Therefore, we have to develop a better understanding of the political challenges and economic issues raised by big data as well as computer tools' modes of operation. This knowledge will help interpret results, especially if we bear in mind both digital data's limitations and its great capacity to represent information.

We chose to follow other researchers in elaborating a critique of purely big data-driven studies on video game communities and practices. Yet, such methodologies' scientific advantages are undeniable. Tracking data on in-game player behaviour provides information that is inaccessible otherwise (even large-scale surveys generate different types of results based on players' own interpretation of their gaming practices). For instance, automated data collection gives us an opportunity to know exactly which functionalities players select (clicks), how avatars move, how frequently they chat, what their gear preferences are, etc. That information can then be correlated with results obtained through different methods, even if it is most often linked with other results generated by computational data. Big data tools primarily operate on the basis of correlation. Their goal is not look at causation, to understand the *why* of a phenomenon. Rather, correlations seek to explain *how* things work together, and, in so doing, identify patterns that should, in all probability, repeat themselves. In addition to this objective of predictability (Mayer-Schönberger & Cukier, 2014), it is worth noting that the tools are often portrayed as helpful in decision-making. Even if explanations can be formulated on the basis of the results produced by big data, the inherent objective of these tools is not to explain, but rather to detect patterns, highlight correlations, and predict phenomena. This form of knowledge is certainly valid, as it can play a role in specific research settings and answer certain types of questions. However, big data cannot answer just any question: it can tell us *what* happens, *how* this process unfolds, and how it will probably unfold in the future, but it cannot explain the *reason why* something is happening. Therefore, mixed methods incorporating qualitative data can help in articulating explanations that go further than description, correlation, and prediction. While social sciences researchers mainly focus on explaining phenomena, they need to adjust their conception of knowledge in the context of big data-driven studies, among other beliefs and habits we discussed above. By way of illustration, when analyzing our data, we could not *explain why* so few male players selected female avatars;

however, we could *predict* what type of players would probably choose female avatars upon creating a new account.

Also, the very fact that data is considered as the basic unit of meaning conveys an underlying conception of science. According to Schroeder (2014), data has three characteristics: 1) it is linked to a research object; 2) its existence predates the analysis process; 3) it is the smallest unit of the analysis. Schroeder notes that "[t]his definition of data has implications for how advance [sic] in social science can be gauged, and presumes a realist and pragmatist epistemology (Hacking, 1983) because the definition requires that there is an object 'out there' (realism) about which more useful or powerful knowledge has been gained (pragmatism)" (2014). Schroeder argues that this constitutes "deterministic knowledge" used to describe human behaviour and that this form of knowledge already implies a certain idea of science.

Finally, the fact that the production of meaning in big data processes are becoming increasingly autonomous, as if machines could automatically perform the 'interpretation' of data without human intervention, is an important epistemological issue. To interpret a sign is to give meaning to it: knowledge is based on our interpretation of the world. While computational technologies may effectively produce some knowledge, it is still necessary that results be kept in check through human intervention (by experts, for instance), that is to say that results and the knowledge they produce be affected with a particular value. This process depends on one's interpretation and judgement. Yet, at a time when, according to Anderson (2008), data and tools give "answers" to questions that were not even asked in the first place, we can ask ourselves what the new role of researchers in social sciences is and, more specifically, how their work might have to become more 'dependent' on big data results. As shown by Kitchen (2014), who successfully rises to the challenge of knowledge production through data-driven science, a consistent critical outlook on knowledge production appears to be more essential than ever.

According to Michel Foucault (1978), critique is exactly that: it is asking questions in a certain manner, it is a practice, "a way of thinking, speaking, and acting" (p. 36, our translation). The philosopher explains: "I seek to place myself outside the culture to which we belong, and to analyze its formal conditions in order to make a critique of them, not in such a way as to diminish its values, but to see how it actually constituted itself" (Foucault, 1967, p. 605, our translation). This epistemological problem also becomes an ethical problem: what is the 'just' stance that researchers must take while using big data tools?

Ethical problems

While Foucault's critique both elaborates and leads to the "art of self-governing" or the "art of not being governed that much" ("*l'art de n'être pas tellement gouverné*", Foucault 1978: 38, our translation), ethics is simply the art of self-government and acts as a guide for our relationships with/in the world. The gathering of data itself presents the first ethical problem on which researchers using big data must reflect critically. In this era of digital surveillance (Andrejevic 2007; Bennett, Haggerty, Lyon & Steeves 2014; Lyon 2015), protecting privacy becomes a major social issue; the phenomenon of big data challenges the established notions of 'private' and 'public'. These questions about privacy and data protection have been raised by many researchers such as Antonio Cassili (2013) as well as lawyer Christopher Marsden and the computer scientist Ian Brown in their book *Regulating Code* (2013). Berendt, Büchler and Rockwell (2015) directly ask the question: "Is it research or is it spying"?

Researchers ought to ask questions about what a *legitimate* rather than *legal* way of gathering data could be: "[j]ust because it is accessible does not make it ethical" (Boyd & Crawford 2012). We may be able to work with legally accessed and collected data, but there is a fine line between players' knowledge of their in-game activities being tracked with their consent, and their ignorance of the phenomenon, its magnitude, and its challenges. While surveillance on video game platforms (computers, connected consoles, smartphones, digital distribution platforms, etc.) seems to be increasingly banalized and in-game reward systems become a surveillance tactic in itself (Surveillance & Society, 2014), we had to question ourselves on the ethical limitations of our data gathering methods (especially those imposed, in our case, by the *Facebook* platform). In this regard, Davis and Patterson, in their book *Ethics of Big Data* (2012), gave a list of questions that any organization or researcher collecting data should ask themselves:

> - Are people entitled to know how their data is used in the business?
> - Are people entitled to know which organization holds their data and what data in particular it holds?
> - Are people entitled to know how their data is analyzed and how the results of these analyses drive the business?
> - Are people entitled to know who within the organization has access to the data? [etc.].

Similarly, Zwitter (2014) talks about "informed consent" and Chessell & al. (2015) problematize the choice given to the person being tracked: "What are the choices given to an affected party? Do they know they are making a choice? Do they really understand what they are agreeing to? Do they really have an opportunity to decline? What alternatives are offered? [etc.]". In other words, any research participant should be able to withdraw at any moment, but how does this work in the context of big data? Do they have any choice now that traceability seems inevitable in the case of online or "connected" games, and especially considering that games are increasingly connected? Although research methods are now being developed in the context of big data, ethical considerations remain crucial to researchers in that they circumscribe their research in line with subjects' best interest and respect. First and foremost, participants in a study must be advised in order to choose whether to give *informed* consent or possibly withdraw from the study.

In our case, the ethical dilemma was extensively discussed both among team members and while applying for our institution's code of research ethics certificate. Jointly with the company, we decided to add a notice within the game's terms and conditions section, regarding the fact that in-game research was being performed. The notice also mentioned that these terms and conditions would be sent to all previously registered users again. Considering that in-game information - that is to say data collected by *Vandal Games* - was not sensitive or detrimental to players' privacy or identity, we decided to use that data and we identified players by their pseudonyms only. We also took factual information compiled by Facebook analytics into account (country and gender displayed) but we did not link that information back to users' individual accounts. In this way, we eliminated all information relative to players' personal *Facebook* accounts, even if these are (too) often publically available. It would have been both possible and legal to access a range of information about the identity of players, and that type of information could have been useful within the context of our inquiry into practices, identity, and socialization. Yet, we considered that accessing such data would be *illegitimate* and we took on reflexions and questions asked by researchers preceding us:

> We are at constant risk of violating participants' rights and our own responsibilities in the conduct of ethical research – because legal rights do not trump ethical ones. […] When players press an "I agree" button, we discovered from a pilot study (Chee & de Castell, 2010), they are not aware of the ways their information is used, shared, or made available to researchers. Informants stated that if they had been aware of these uses being made of their game-based information, they would not have agreed to them. [...] How are games researchers dealing ethically with the requirement of informed consent?
>
> de Castell & al., p. 138 (2012)

If researchers must address these ethical questions regarding the individuals who take part in their study, research tools themselves must be interrogated as well, because they convey certain values: "Big Data is mistakenly framed as morally neutral or having benefits that overweigh any costs" (Martin, 2015). The underlying logic of massive data gathering connotes a specific representation of the world based on quantification, which is not exempt of ethical problems. Moreover, some scholars have addressed "algorithmic governmentality" (Rouvroy & Berns, 2013) or reiterated that massive data gathering is a form of social control that "cannot be separated from neoliberal approaches to governance" (Langlois, Redden & Elmer, 2015, p. 6). The scope of all of these ethical, political, and economic challenges and issues is much wider than that of studies of player behaviour. Nevertheless, many contributors to the book *Compromised Data* (Langlois, Redden & Elmer, 2015) ask researchers using big data tools to take time to reflect more deeply on those issues.

This is especially relevant since the results produced by big data tools not only affect both game designers' decisions and game development, but also impact players themselves in doing so. Indeed, big data's predictive results are sent back to players as they are exploited in games, and this affects players' gaming experience. The possibility to influence player behaviour entails an important responsibility for researchers to the extent that they may actually prompt the behaviour they predicted. This phenomenon may be called the self-fulfilling prophecy (Merton, 1948) and it ties in with Baudrillard's concept of hyperrealism (1981), which describes the way in which representations come to produce reality. Therefore, in conducting our study and reflecting on the "art of self-governing", we were pushed to question our own relationship to big data tools and the results generated by their use.

Conclusion

This article did not aim to present results from the study we conducted, but rather to elaborate on our research experience and especially to shine a light on the problems and questions that we faced as Semiotics and Communications scholars within the field of game studies. These problems included coordinating the research team, adapting to a business environment, learning about computer tools and their logic, understanding the limits of data and results' validity, as well as identifying ethical challenges and constraints. Thus, we have learned to adapt our research methods to this new approach.

Even though we have been confronted with new ways of working and understanding knowledge, the exploration of a computer world in which considerable amounts of data are produced and processed has also fascinated us. Research that uses big data tools requires a better assessment of its epistemological conditions and challenges, and researchers must keep in mind the ethical risks that it presents. Nonetheless, this form of research has its advantages,

as it reveals information that would otherwise be impossible to access. Considering it as what it is and taking the needed precautions in doing so, it appears that digital data offers information that can serve scientific purposes, particularly in studying player behaviour. Given the sheer amount of information big data permits us to comprehend, generalizing behavioural patterns gives us an insight into player communities and gaming experiences that could not have been obtained otherwise.

Indeed, data not only gives us specific information about the game, but the possibility of cross-referencing different types of information also allows for a more intimate understanding of communities. Considered not so much as a unique research method but rather as one method among many others, research with big data tools offers scientific advantages particularly when the research team is mixed, bringing computer and social sciences scholars together. Nevertheless, we must keep in mind at all times that our fascination with these increasingly efficient tools could blind us: we ought to stay critical, "analyze our limits, and reflect on them" (Foucault, 1984, p. 1383, our translation).

References

Anderson, C. (2008, June 23rd). The end of theory: The data deluge makes scientific method obsolete. *Wired*.

Andrejevic M. (2014). The big data divide. *International Journal of Communication*, 8, 1673-1689.

Andrejevic, M. (2007). *iSpy: Surveillance and power in the interactive era*. Lawrence: University Press of Kansas.

Bainbridge, W. S. (2007, July 27). The scientific research potential of virtual worlds. *Science,* 317(5837), 472-476.

Bennett, C. J., Haggerty, K. D., Lyon, D. & Steeves, V. (2014). *Transparent lives: Surveillance in Canada*. Edmonton: Athabasca University Press.

Berendt, B., Büchler, M. & Rockwell, G. (2015). Is it research or is it spying? Thinking-through ethics in big data AI and other knowledge sciences. *Künstliche Intelligenz*, 29(2), 223-232.

Blizzard (2016). *World of Warcraft Forums*. Retrieved from http://us.battle.net/wow/en/forum/ [accessed on January 14th, 2016].

Boellstorff, T., Nardi, B., Pearce, C., & Taylor, T. L. (2012). *Ethnography and virtual worlds: A handbook of method.* New Jersey: Princeton University Press.

Boellstorff, T. (2013, October). Making big data, in theory. *First Monday*, 18(10).

Bollier, D. (2010). *The promise and perils of big data.* Washington, DC: The Aspen institute.

Boyd, D. & Crawford K. (2012). Critical questions for big data. *Information, Communication & Society*, 15(5), 662-679.

Brown, I. & Marsden, C. (2013). *Regulating code: Good governance and better regulation in the information age*. Cambridge: MIT Press.

Buchanan, E. & Markham, A. (2012). *Ethical decision-making and internet research: Recommendations from the AoIR ethics working committee (Version 2.0)*. Association of Internet Researchers.

Busch, L. (2014). A dozen ways to get lost in translation: Inherent challenges in large-scale data sets. *International Journal of Communication*, 8, 1727-1744.

Cassili, A. (2013). Contre l'hypothèse de la 'fin de la vie privée': La négociation de la privacy dans les médias sociaux. *Revue Française des Sciences de L'information et de la Communication*, 3.

Cassin, B. (2007). *Google-moi: La deuxième mission de l'Amérique*. Paris: Albin Michel.

Charles, D., & Black, M. (2004). Dynamic player modelling: A framework for playercentric digital games. In *Proceedings of CGAIDE 2004, 5th international conference on computer games: Artificial intelligence, design and education*, Microsoft Campus, Reading, UK.

Chessell, M. (2014). *TCG study report: Ethics for big data and analytics.* IBM.

Chessell, M., Sivakumar, G., Wolfson, D., Hogg, K., & Harishankar, R. (2015). *Common Information Models for an Open, Analytical, and Agile World*. IBM Press.

Consalvo, M. & Ess, C. (2011). *The handbook of internet studies*, New Jersey: Wiley-Blackwell.

Crawford, K., Milner, K. & Gray, M. L. (2014). Critiquing big data: Politics, ethics, epistemology. *International Journal of Communication*, 8, 1663-1672.

Davis, K. & Patterson, D. (2012). *The Ethics of big data: Balancing risk and innovation.* Cambridge: O'Reilly.

de Castell, S., Jensen, J., Taylor, N. & Weiler, M. (2012). Theoretical and methodological challenges (and opportunities) in virtual worlds research. *Proceedings of the International Conference on the Foundations of Digital Games*. ACM.

Drachen, A., Canossa, A. & Yannakakis, G. N. (2009, 7-10 Septembre). Player modeling using self-organization in Tomb Raider: Underworld. *Computational Intelligence and Games 2009*. Actes du colloque, Milano.

Drachen, A., Sifa, R., Bauckhage, C., & Thurau, C. (2012). Guns, swords and data: Clustering of player behavior in computer games in the wild. I*n Proceedings of IEEE computational intelligence in games*. Granada, Spain.

Ducheneaut, N., & Moore, R. J. (2004). The social side of gaming: A study of interaction patterns in a massively multiplayer online game. *In Proceedings of the 2004 ACM conference on computer supported cooperative work*. Chicago, Ill.

Ducheneaut, N., Yee, N., Nickell, E., & Moore, R. J. (2006). Building an MMO with mass appeal: A look at gameplay in World of Warcraft. *Games and Culture*, *1*(4), 281–317.

Ducheneaut, N., Yee, N., Nickell, E. & Moore, R. J. (2007). The life and death of online gaming communities: A look at guilds in World of Warcraft. In *Proceedings of the SIGCHI conference on Human factors in computing systems,* p. 839–848. ACM.

El-Nasr, M., Drachen, A., & Canossa, A. (2013). *Game analytics: Maximizing the value of player data*, Springer, Londres.

Ellul, J. (1954). *La technique ou l'enjeu du siècle*. Paris: Armand Colin.

Euclid. (1896), *Data*, in *Euclidis Opera Omnia*, Vol. 6, H. Menge. Leipzig: Teubner.

Foucault, M. (1967). Qui êtes-vous professeur Foucault. Dans D. Defert, F. Ewald et J. Lagrange (dir.). *Dits et écrits I, 1954-1969,* p. 601- 619. Paris: Gallimard.

---------- (2001 [1971]). *Dits et écrits I, 1954-1975*, Paris, Gallimard, Quarto.

--------- (1976). *La volonté de savoir*, Histoire de la sexualité. Paris, Gallimard.

--------- (1978 [1990]). Qu'est-ce que la critique? Critique et *Aufklärung*. *Bulletin de la Société Française de Philosophie*, *84*(2), 35-64.

Gitleman, L. (2013). *Raw data is an oxymoron*. Cambridge: MIT Press.

Hacker, I. (1983). *Representing and intervening*. Cambridge: Cambridge University Press.

Hoobler, N., Humphreys, G., & Agrawala, M. (2004). Visualizing competitive behaviors in multi-user virtual environments. In *Proceedings of the conference on visualization.* Los Alamitos: IEEE.

Hopson, J., Hullett, K., Nagappan, N. & Schuh, E. (2011, 21-28 Mai). Data analytics for game development: NIER track. *Software Engineering (ICSE), 2011 33rd International Conference*. Actes du colloque, Honolulu.

Kennerly, D. (2003). Better game design through data mining. *Gamasutra*.

Isbister, K. & Schaffer, N. (2008). *Game usability: Advancing the player experience*. San Francisco: Morgan Kaufman.

Kim, J. H., Gunn, D. V., Pagulayan, R. J., Phillips, B. C., Schuh, E. & Wixon, D. (2008, 5-10 avril). Tracking real-time user experience (TRUE): A comprehensive instrumentation solution for complex systems. *Conference on human factors in computing systems*. Actes du colloque, Florence.

Lazer, D., Kennedy, R., King, G. & Vespignani, A. (2014). The parable Google flu: Traps in big data analysis. *Science*, 343, 1203-1205.

Leibniz, G.W. (1698 [1987]). Lettre au Père Verjus: Fin de l'année 1698. *Discours sur la théologie naturelle des Chinois*. Paris: L'Herne.

Lewis, C. & Wardrip-Fruin, N. (2010, 19-21 Juin). Mining game statistics from web services: A World of Warcraft armory case study. *Fifth International Conference on the Foundations of Digital Games*. Actes du colloque, Monterey.

Loudon, M., Maurer, B., Norton-Ford, J., Fricker, M. & Blumenstock, J. (2013). Big data in ICT4D: What can we learn from prepaid mobile airtime transactions? *Proceedings of ICTD 2013: Sixth International Conference on Information and Communication Technologies and Development.* Cape Town, South Africa.

Lyon, D. (2015). *Surveillance after Snowden*. Cambridge: Polity Press.
Magerko, B. & Medler, B. (2011). Analytics of play: Using information visualization and gameplay practices for visualizing video game data. *Parsons Journal For Information Mapping*, 3(1).
Martin, K. E. (2015). Ethical issues in the big data industry. *MIS Quaterly Executive*, 14(2), 67-85.
Medler, B. (2011). Player dossiers: Analyzing gameplay data as a reward. *The International Journal of Computer Game Research*, 11 (1).
Missura, O., & Gärtner, T (2009). Player modeling for intelligent difficulty adjustment. In *Proceedings of the 12th international conference on discovery science*, DC'09, Berlin.
Moura, D., Seif El-Nasr, M. & Shaw, C. D. (2011). Visualizing and understanding players'
behavior in video games: Discovering patterns and supporting aggregation and comparison. *Proceedings of the 2011 ACM SIGGRAPH symposium on video games (Sandbox '11),* 11–15. New York.
Mumford, L. (1967). *The myth of the machine (vol. 1): Technics and human development*. San Diego: Harcourt Brace Jovanovich.
Mumford, L. (1970). *The Myth of the Machine* (vol. 2): *The Pentagon of Power*. San Diego: Harcourt Brace Jovanovich.
Nardi, B., Pearce, C. & Taylor. T. L. (2012). *Ethnography and virtual worlds*. Princeton University Press.
Poor, N. (2015). *What MMO communities don't do: A longitudinal study of guilds and character leveling, or not.* Palo Alto: AAAI.
Pushmann, C. & Burgess, J. (2014). Metaphors of big data. *International Journal of Communication*, 8, 1690-1709.
Reeve, C. D. C. (2004), *Plato: The republic.* Indianapolis: Hackett.
Rouvroy, A., & Berns, T. (2013). Gouvernementalité algorithmique et perspectives d'émancipation". *Réseaux*, *177*(1), 163–196.
Schroeder, R. (2014). Big data and the brave new world of social media research. *Big Data & Society*, *1*(2).
Simondon, G. (1958). *Du mode d'existence des objets techniques*. Paris: Éditions Aubier-Montaigne.
Stiegler, B. (1994). *La technique et le temps: La faute d'Épiméthée* (tome 1). Paris: Galilée.
--------- (1996). *La technique et le temps*: *La désorientation* (tome 2). Paris: Galilée.
--------- (2001). *La technique et le temps: Le temps du cinéma et la question du mal-être* (tome 3). Paris: Galilée.
Thurau, C. & Bauckhage, C. (2010). Analyzing the evolution of social groups in World of Warcraft. *Proceedings of the international conference on Computational Intelligence and Games*, IEEE, CIG'10. Copenhagen.
Tychsen, A. (2008). Crafting user experience via game metrics analysis. *Proceedings of the Workshop 'Research Goals and Strategies for Studying User Experience and Emotion' at the 5th Nordic Conference on Human-computer interaction: Building bridges (NordiCHI)*. Lund, Sweden.
Vandal Games. (2012-2016). *Big Story Little Heroes*. [Jeu vidéo]. Montréal: Vandal Games.
Weinberger, D. (2014). *Too big to know, rethinking knowledge now that the facts aren't the facts, experts are everywhere, and the smartest person in the room is the room*. New York: Basic Books.
Whitson, J. R. (2013). "Gaming the quantified self". *Surveillance & Society*, *11*(1/2), 163–176.
Williams, D. (2010). *The promises and perils of large-scale data extraction*. Chicago: McArthur Foundation.

Williams, D., Ducheneaut, N., Xiong, L., Zhang, Y., Yee, N., & Nickell, E. (2006). From tree house to barracks: The social life of guilds in World of Warcraft. *Games and Culture*, *1*(4), 338–361.

Williams, D., Yee, N. & Caplan, S. E. (2008). Who plays, how much, and why? Debunking the stereotypical gamer profile. *Journal of Computer-Mediated Communication*, *13*(4), 993–1018.

Williams, D., Consalvo, M., Caplan, S., & Yee, N. (2009). Looking for gender (LFG): Gender roles and behaviors among online gamers. *Journal of Communication*, 59, 700–725.

Wood, R. T., Griffiths, M. D. & Eatough, V. (2004). Online data collection from video game players: Methodological issues. *CyberPsychology & Behavior*, 7(5), 511–518.

Zwitter, A. (2014). Big data ethics. *Big Data & Society*, 1(2).

---

[1] http://www.sshrc-crsh.gc.ca/funding-financement/umbrella_programs-programme_cadre/insight-savoir-fra.aspx