Journal of
Virtual Worlds Research
jvwresearch.org ISSN: 1941-8477

Volume 3, Number 3
The Researcher's Toolbox, Part II
May 2011

Editor-in-Chief Jeremiah Spence

Image Art ©2007-2011 ~enchanted-stock
http://fav.me/dwquz0

Technical Staff John Brengle
Betsy Campbell
Sil Emerson

Journal of Virtual Worlds Research
jvwresearch.org ISSN: 1941-8477

Volume 3, Number 3

The Researcher's Toolbox, Part II

May 2011

Collecting conversations: three approaches to obtaining user-to-user communications data from virtual environments

Mika Lehdonvirta
The University of Tokyo, Japan

Vili Lehdonvirta
Helsinki Institute for Information Technology, Finland

Akira Baba
The University of Tokyo, Japan

Abstract

Transcripts of conversations are a valuable research resource in social sciences and can be used to make inferences about subjects' behavior and intentions. Large-scale communications records can be coded and analyzed statistically for generalizable results. Virtual environments are a good place to gather communications records, because they exhibit a wide variety of subject behaviors However, compared to traditional channels such as forums and chat rooms, virtual environments can be more challenging to obtain data from. In this article, we describe three approaches to collecting user-to-user communications data from virtual environments: requesting back-end records from the operator of the environment, recruiting "data donors" among the users, and setting up researchers' own "listening posts". The data collection approaches are evaluated empirically in *Uncharted Waters Online*, a Japanese massively-multiplayer game. Avatar gender ratio is used as a diagnostic variable to compare the representativeness of the resulting data sets. Both data donors and listening posts yielded data with a gender ratio that corresponds to the back-end records, but the back-end gender ratios differed significantly between two different servers. We conclude that all three approaches can be statistically viable: the choice of method depends more on desired sampling scope and on practical factors such as resources and timetable; but when defining a sampling frame, it cannot be assumed that one server is necessarily representative of the whole platform.

Keywords: chat log; content analysis; lurking; methodology; online game; research ethics

Collecting conversations: three approaches to obtaining user-to-user communications data from virtual environments

There is growing interest in the use of virtual environments as platforms for social scientific research. An often cited advantage of virtual environments for research purposes is the ability to accurately record large amounts of data on participants' actions within the environment (Blascovich et al., 2002; Yee and Bailenson, 2008). However, not all behaviors can be inferred from actions alone. The behavioral significance of an action often depends on the intent and meaning attached to it by the persons involved. For example, it may be impossible to distinguish between pro social behavior such as gift-giving and anti-social behavior such as extortion by observing the mere flow of goods and money.

So far only a few studies have used activity records from a large-scale virtual world for research purposes (Castronova et al., 2009; Duchenaut et al., 2006; Harris et al., 2009; Lehtiniemi, 2009). Most empirical studies pertaining to virtual worlds continue to collect data through more traditional methods, such as surveys, interviews and ethnographies (e.g., Hussain and Griffiths, 2008; Johnson and Sihvonen, 2009; Steinkuehler and Williams, 2006). These methods examine the world through participants' eyes and therefore avoid the problem of how to interpret actions as behaviors But while doing so, they also give up on the supposed advantages of virtual worlds as research platforms, and particularly on the notion of large-scale directly recorded data that promises accuracy and generalizability.

In this article, we focus on a type of data that can be described as falling somewhere between the aforementioned direct but dumb action records and deep but mediated interviews and ethnographies: records of user user-to-user communications, especially typed chat conversations. As with action records, digital media in theory allows for direct, accurate and comprehensive recording of user-to-user communications. But in contrast to action records and similar to interviews and ethnographies, communications data pertains to the social reality of the participants, and is thus useful for studies seeking to understand and explain various aspects of participants' behavior (Leuski and Lavrenko, 2006). Communications data represents something of an underutilized resource in virtual worlds research, and the aim of this article is to help researchers tap it.

Communications data obtained through the methods described in this article is typically analyzed further using various content analysis methods (Neuendorf, 2002). In particular, communications data in the form of textual records can be coded into a quantitative form by representing words, topics, persons and other elements with numbers

according to a pre-defined coding scheme. The encoded data can then be analyzed using statistical methods to discover patterns or examine hypotheses. Collecting sufficiently large samples of conversations in offline settings is often laborious. This research approach thus utilizes the special potential of virtual environments as platforms for collecting large amounts of data for social scientific research.

In computer-mediated communication research, user-to-user communications records have been used for research purposes for over two decades (Cheseboro and Bonsall, 1989). They are typically obtained through "lurking": unobtrusive observation on public forums such as chat rooms and newsgroups, a method which is well documented in the literature of Internet research methodology (Kozinets, 2006). However, massively-multiplayer games (MMO) and other avatar-driven online environments have important differences to more traditional online hangouts. For instance, the way in which they mimic physical space recalls some of the challenges that unobtrusive observation faces in non-mediated natural settings. There is a gap in the literature when it comes to collecting communications data from virtual environments. Yet as virtual environments permit a greater variety of behaviors than most traditional computer-mediated communication channels, data from them is valuable for research.

In this article, we describe three different approaches to obtaining large-scale user-to-user communications data from virtual environments, and discuss the challenges associated with each approach, including ethical issues. We then describe a series of data collection efforts where these approaches are applied in practice, illustrating our particular solutions to the challenges. The data collection takes place in *Uncharted Waters Online* (UWO), a Japanese MMO launched in 2005 by KOEI Corporation. The data sets obtained through the different methods are then compared against each other, yielding insights regarding the relative merits of each approach. Finally, we draw some conclusions in the form of suggested practices in collecting user-to-user communications records from virtual environments.

Approaches to collecting communications data

As methods of communication vary from one multi-user environment to another, what exactly is meant by user-to-user communication data is different in each environment. Here the primary focus is on chat data: lines of text written by participants, each line indicating which participant wrote it. The lines are typically ordered chronologically and may include time stamps. Chat data is also often organized into "channels": each line indicates which channel, or group of potential recipients, it was directed to. In this section, we outline three

different sources from which chat data can be obtained: back-end logs, "listening posts" set up by the researchers, and "data donors" supplying researchers with their own chat logs.

Back-end data

In a typical multi-user online environment, all communications between participants pass through a server maintained by the operator of the service. This operator, typically a commercial company such as a game publisher, is thus in a position to collect a complete record of all the chat conversations taking place on the platform. If researchers can obtain access to this data, it is obviously an excellent resource. In our experience, obtaining access typically involves knowing someone in the company, presenting them with a research proposal that highlights the benefits of the study to the company, and drafting a legal agreement that authorizes the use of the data. A good place to meet MMO operators is at industry events and conferences. The main benefits that good research can offer to an operator are design insights, customer intelligence, and publicity. The operator's legal team will probably propose an agreement that gives the company veto over publication, which is risky for the researcher.

Once a data set is obtained, its sheer size may present challenges for subsequent handling and analysis. For example, one of us once received a file of 540 gigabytes from an MMO operator. But for most purposes, there is no need to handle the complete set: a suitably selected sample will yield the same results with less processing.

In practice, researchers are not often able to gain access to extensive data from commercial operators. This can be due to a combination of costs, trade secrets, organizational inertia, and real and imagined legal issues. Through our collaboration with the KOEI Corporation, we were able to gain access to a complete database of the attributes of every avatar in UWO, but not to any chat logs, which the company considered too sensitive. Even when successful, it should be noted that a data request can be an arduous undertaking. The process that resulted in the transfer of the 540 gigabyte file mentioned above took over 12 months from the day the chief executive officer of the company agreed to the idea. Scholars can be opportunistic about their research, studying whatever platform they find easiest to obtain data on from a benevolent operator. But while such an approach is fine for graduate students, it is not a very good basis for a long-term research strategy. Therefore, it is important to consider data collection methods that do not require active assistance from the operator of the platform.

Listening posts and data donors

A basic data collection approach for any social scientist working in natural environments (as opposed to experimental settings, interviews or surveys) is observation: observing people and recording what they do or say according to a pre-determined scheme (Creswell, 2003). This unobtrusive observation can be contrasted with participant observation, a method often used by ethnographers, which involves finding a role in the community under scrutiny and reporting on its practices from the "inside". While participant-observers may be able to learn and understand things that are not intelligible to outside observers, they risk influencing the phenomenon they are studying through their visible participation.

In the context of computer-mediated communication, unobtrusive observation is informally known as "lurking" and has already been used as a data collection method for two decades. In computer-mediated channels, such as chat rooms, mailing lists, newsgroups, and discussion forums, it is particularly easy to collect large amounts of data without any awareness from the subjects. Unobtrusive observation can likewise be applied in MMOs and other avatar-driven online environments. For example, Yee and Bailenson (2008) describe a method for capturing data from the virtual environment *Second Life,* through the eyes of an avatar. However, the way in which these environments mimic physical space recalls some of the challenges that unobtrusive observation faces in non-mediated settings. One such challenge is that avatars' perceptions are often programmed to be limited to their immediate surroundings and to the chat channels that the avatar is a member of. Depending on what kind of a sample is sought, it may therefore be necessary for the researcher to establish a large number of avatars as "listening posts" in various locations and channels to get sufficient coverage. Another challenge is that the observer's avatars are often visible to the subjects, raising concerns about the influence of the observer's presence on the subjects' behavior

An alternative to establishing artificial listening posts is to ask the participants themselves to record and provide chat data to the researchers. This can be done by, for example, posting a request on a forum frequented by the users of the environment. In many online environments, the client program has a built-in chat logging feature. In some others, logging can be achieved using third-party software. Chat logs obtained from "data donors" in this way typically consist of text appearing in chat channels of which the donor is a member during the times that the donor is online. Collection is therefore limited to certain channels and time intervals, but the data will include the donors' private conversations, which are not otherwise observable.

The potential influence of researchers' presence on subjects' behavior may be avoided if researchers ask donors to provide past logs which were recorded without any knowledge that they will be used for research purposes in the future. However, this collection method may introduce a potentially more serious problem: the risk of data donors censoring, altering, or otherwise tampering with the conversation logs before submitting them to the researchers, introducing social desirability and other biases. Data donors are also largely self-selected, which is traditionally a red flag in research involving statistical reasoning, because it may introduce bias to the sample. On the other hand, data obtained from donors will also include conversations by other participants, who are not self-selected.

Representativeness

When chat data is gathered for the purpose of coding and subsequent statistical analysis, as opposed to pure qualitative analysis, the ultimate objective is usually to generalize the findings to some larger population. Sometimes this statistical population is the active user base of the online environment we are studying. For example, our research question might ask us to find out how many percent of the active users of the environment remember to say thanks after receiving advice. In this case, we would need chat data from a sample of users that is representative of all the active users. The best way to ensure representativeness is to select the sample users at random from the whole population. This is possible if we have access to the complete chat records of the whole platform. But if we are relying on data from listening posts or data donors, obtaining a broadly representative sample is challenging.

We should obviously take all available precautions to ensure that our sample is as random as possible: distribute listening posts evenly, strive to obtain submissions from different kinds of data donors, and collect data over a reasonably long period of time. We should also consider whether data needs to be collected from multiple servers, or whether one server can adequately represent the whole platform. But despite these precautions, some easily identifiable biases are likely to remain. For example, no matter how evenly the listening posts and data donors are distributed in virtual and social space, they will be gathering clusters of conversations from their immediate vicinities, as opposed to truly randomly distributed conversations. The researchers must acknowledge these biases and assess what kind of an impact they may have on the validity of the results. For instance, it might be plausible to suggest that team players differ in their thanking propensity from solo players. If the sample clusters are biased towards team players, our ability to draw valid inferences from the sample to the whole user base is diminished. The good news is that for

many variables on which sampling bias occurs, such as whether the conversations took place in virtual town A or the neighboring town B, a plausible link to the behavior under scrutiny can scarcely be suggested. It is less important that the sample is truly random: what is important is that the sample is representative when it comes to variables that can reasonably be expected to influence the outcome.

In studies that use online environments as platforms for social scientific research, as opposed to being interested in the properties of the online environment itself, it is typical that the statistical population to which the findings are generalized is actually something different from the user base of the service. For example, our research question might ask us to examine whether there is a difference in thanking propensity between males and females. In this case, any sampling biases inside the platform, such as a bias towards team players, should not have impact on the validity of our conclusions, *unless* the biased variable has an interaction effect with gender. In other words, if being a team player influences the thanking propensity of both male and female users in the same way, then the sampling bias does not matter. But if there is a reason to expect that being a team player might have a different effect on males and females, then the bias diminishes the validity of the findings. On the other hand, whether or not the gender distribution of the sample is representative of the whole user base is irrelevant, because we are not aiming to make statements about the user base.

In any case, a healthy amount of reservation is necessary when generalizing conclusions from online environments to other settings. The specifics of the particular environment where the interaction takes place are sure to have an impact on the behavior, and in the worst case, an interaction effect with an explaining variable.

Sample size and reliability

How large a sample is necessary? It depends on the aims of the research. A typical aim of statistical tests conducted on the sample is to demonstrate that a hypothesis regarding the whole population is true with a certain probability, usually 95, 99 or 99.9 percent. Assuming that we are able to obtain a random sample, the sample size that will yield this probability can be calculated from other parameters (see e.g., Hair et al., 2006). But there are two issues in applying this approach to estimating chat data sample size. The first is that unless we have access to back-end chat records, our sample is probably not random. The second is that the calculations referred to above actually yield the required number of *cases*, not lines of chat text. For example, if we were trying to find out how many percent of the user population remember to say thanks after receiving advice, there is a formula for calculating how many

individuals (i.e., cases) we need to observe, but no a priori means of determining how many lines of text we need to gather to obtain those cases (cases are obtained from chat data through coding, for example).

An alternative to a priori estimations of required sample size is the notion of sampling saturation (Glaser and Strauss, 2006). In qualitative data collection (e.g., interviews), a sampling process that aims for saturation involves adding data to the sample incrementally, until additional data ceases to produce any additional categories or insights; that is, until the sample is saturated. In the context of quantitative and quasi-quantitative data such as user-to-user communication records, the process involves adding quasi-random data to the sample until the distributions of important variables stabilize For example, as we add more lines of text to the sample, the proportion of male and female avatars converges towards what is the true gender ratio for the service, assuming our sampling method is not biased on this variable. Once additional data no longer significantly changes important distributions, we may conclude that the sample is adequately large. Regardless of which sampling method is used, if the sample size is too small, even true random sampling will yield results that are not reproduced if the study is repeated; that is, the study will suffer from a lack of reliability.

Ethical considerations in collecting communications data

Since the research methods discussed in this article pertain to human subjects, there is a special need to consider possible ethical issues that the methods may give rise to. Ethical considerations in the use of human subjects in research have been a topic of intense deliberation and codifying for several decades (Shamoo and Khin-Maung-Gyi, 2002; Shamoo and Resnik, 2003). However, in the areas of Internet-enabled research and online data collection, many questions remain far from settled (Bos et al., 2009; Bruckman, 2006; McKee and Porter, 2009). This is partly because the research methods and research environments are relatively new and still constantly changing and developing, and also because the research involves many disciplines and tends to extend across jurisdictions. The purpose of this section is to offer a brief introduction to some of the ethical issues relating to the data collection practices described in this article, and provide some ideas on dealing with them in practice. For more thorough discussions of the ethics of online data collection, see Bruckman (2002), Hudson and Bruckman (2005) and Kleinberg (2007). For a discussion of the ethics of MMO related research, see McKee and Porter (2009).

Various declarations and codes of conduct that govern researchers' practice recognize that researchers must pay special attention to human dignity, privacy, and the confidentiality

of communications (Shamoo and Resnik, 2003). Privacy and non-interference with communications are moreover human rights prescribed in the Universal Declaration of Human Rights and protected by law in many jurisdictions. Researchers collecting communications data from online environments must thus be careful not to infringe upon these rights.

One way of fulfilling this duty is to obtain consent from each subject for recording communications to which they are party. Obtaining consent is standard procedure in many fields, such as medical sciences, where human subjects research is common (Shamoo and Khin-Maung-Gyi, 2002). However, for the online data collection methods described in this article, obtaining consent presents three challenges. Firstly, the methods are intended to capture data pertaining to a large number of subjects, making it unfeasible to request consent from each subject personally. Legal consent could feasibly be obtained as part of the environments' terms of service with the help of the operator company. However, it is not clear how informed such "automatic" consent would be, so that while legal requirements might be satisfied, the ethical duty might not be. Secondly, the methods aim to be unobtrusive, so as to capture data unaffected by the researchers' presence. A prominent request for consent could compromise unobtrusiveness. Finally, the methods aim to capture communications, and communications always involve at least two parties. Consent is not helpful unless it can be obtained from both.

Without consent, it is still possible to collect communications data ethically, but more consideration needs to be given to what sources are appropriate. The basic rule is that private communications are off-limits, but researchers who limit their observations to communications that are intended to be public, such as public speeches and newspaper correspondence, have traditionally been considered on the safe side (Bruckman, 2002). If the same logic is extended to online environments, all communications published on publicly accessible media could be considered public and non-infringing on privacy.

However, there are different expectations of privacy and anonymity pertaining to different modes of online communications (Bruckman, 2002; McKee and Porter, 2009). For example, chat rooms that are in theory publicly accessible may nevertheless be felt to be quite private if in practice the attendance is usually limited to a few individuals. Chat room conversations may also be perceived as more ephemeral than blogs or discussion board conversations, an expectation that can be violated by archival. Consequently, while ethics review boards and codes of conduct often rely on dichotomies such as "public" vs. "private," "published" vs. "unpublished," and "anonymous" vs. "identified", online data sources

frequently require more delicate consideration of social expectations to understand the limits of ethical behavior (Bos et al., 2009; Hudson and Bruckman, 2005).

As long as the data collection method is legally sound, it may be possible to fulfill additional moral duties of protecting privacy and dignity by means of anonymization. Data anonymization involves manipulating a data set in such a way that the people or groups described by the data become unidentifiable, while the research value of the data is preserved. As the data set is subsequently handled by researchers, shared with other research groups, published online or even inadvertently leaked, subjects' privacy is preserved. Anonymizing data already as it is collected is a stronger protection than the standard practice of withholding subjects' identities when the results of the research are published.

True anonymization is not easy. Supposedly anonymous search queries released by AOL and video rental records released by Netflix have been successfully de-anonymized and linked to individual persons (Narayanan and Shmatikov, 2008). In the latter case this was achieved through sophisticated data analysis methods, while in the former case reading the queries and doing some simple detective work was sufficient. Coding methods that reduce the subjects and contents of a data set into numbers or symbols in preparation for statistical analysis can be useful in this respect, since they discard potentially exploitable information that is not needed in the research.

It could be argued that online communications are often anonymous to begin with, as they are frequently conducted under a pseudonym, such as an avatar name. This argument does not stand well, however. In some cases the connection between a pseudonym and a legal name is public knowledge. And in any case, it can be argued that people should be able to expect privacy regardless of which name they are acting under (McKee and Porter, 2009). Data anonymization should thus apply equally to pseudonymous conversations.

Data collection in practice: case Uncharted Waters Online

Uncharted Waters Online involves participants assuming the role of merchants, explorers, and privateers in a 17th century world where sailing is the main means of transport. Players form "companies" similar to guilds in many other MMOs. According to a survey conducted by KOEI in 2006 (N=5898), 87 percent of the game's participants are male and 13 percent are female. All age groups are represented, but 44 percent of the participants are in their twenties and 47 percent in their thirties. We wanted to collect chat data from UWO for a research project on pro-social behavior in online environments. We collected data in two instances, first using the data donor approach, and later setting up our own listening posts. In

this section, we describe how the data collection was carried out in practice, and compare the results of the different approaches.

To collect chat logs from data donors, we posted a short appeal on the game's official forums, requesting players to email us their unmodified chat logs from the past 30 days for research purposes. The request was limited to a selected server, "Euros". By requesting past logs, we sought to obtain data unaffected by the subjects' awareness of being observed. Within one week, three participants contacted us and provided us with their chat logs as requested. We removed the donors' own conversations from the data, and instead used only other peoples' conversations overheard by the donors. This was done for three reasons: to mitigate possible self-selection bias, to mitigate the over-representation of the donors' own conversations in the sample, and to avoid any issues of privacy and bias relating to the use of private conversations. The remaining data represented conversations on public channels and semi-public company channels at various locations in the game world. These conversations amounted to a total of 271,939 lines of text, which is a relatively large chat log in terms of quantity.

Despite having successfully obtained a sizable set of chat data, we were concerned about possible biases due to the relatively small number of data donors. Six months later, the first author planned and executed a second, more ambitious data collection effort using the listening posts approach. The objective was to use a large number of randomly positioned observation points to significantly alleviate sampling bias. As the main research interest was on group behavior, the collection was this time limited to conversations on company channels. The data collection effort is described below.

A roomful of listening posts

Preparations for the actual data collection consisted of creating the avatars used as "listening posts" and positioning them in different company chat channels. 82 accounts to the game were obtained from KOEI. Two avatars were created on each account, one for each of the two servers chosen for the study, "Euros" and "Zephyros". All user-definable avatar attributes were randomized (e.g., gender, appearance, starting location). Avatar names were created using a random fantasy character name creator software. The resulting 164 avatars and their servers and locations were recorded into an Excel file.

Ten days before the data collection was to commence, each character sent an application letter to the head of a randomly selected company on its server. A decision was made to target only companies with 30 or more members in order to exclude inactive

companies and companies consisting mainly of multiple characters belonging to a single player or a small number of players. No two avatars applied for the same company. The application letter consisted of a single phrase chosen at random from a list of 20 phrases commonly used in this situation. UWO companies are generally friendly towards beginners. 77 percent of the application letters resulted in the avatar being admitted to the company.

The actual data collection was performed over two weekends in a computer class at a school. Weekends were chosen because more players are generally online, and the players probably represent a wider variety of backgrounds compared weekdays, when many players are busy with their occupations. The UWO client program was successfully installed and run on 48 computers in the class, so 48 characters could be logged in simultaneously. Two groups of 48 avatars were thus chosen from the pool of available avatars, each group having 24 avatars from each of the two servers. Group A was logged in to the game on the Saturday of the first weekend and the Sunday of the second weekend, while group B was logged in on the Sunday of the first weekend and the Saturday of the second weekend.

When an avatar was logged in, it was immediately directed to join the company's chat channel and to utter a generic greeting. After a few moments, it was then directed to indicate that it was going "AFK" (away from the keyboard) for a while. Actual data logging was commenced after the avatar had been AFK for several hours and therefore hopefully obtained the status of an unobtrusive "lurker" that would not influence conversations on the channel. Logging was performed for approximately 12 hours on each of the four days. All the listening post avatars were also members of a private chat channel created for the purposes of administering the study. This channel was used to insert synchronized "BEGINNING OF DATA" and "END OF DATA" marks in the chat record of each listening post. Only the text falling between these lines was taken forward to analysis.

Comparing the results

We were now in possession of two sets of similar chat data, one obtained through the data donor method and the other through listening posts. In addition, we had access to a back-end database of avatar attributes. We could thus compare the data yielded by each of the methods against each other, and against the baseline provided by the back-end data. The data sets consisted of data from two different servers, so we could furthermore examine the question of whether multi-server sampling is necessary to achieve representativeness.

In terms of raw quantity, the listening post data amounted to a total of 129,595 lines of text, which is less than half of the 271,939 provided by the data donors. This may seem

surprising, given the large number of listening posts employed. But considering the differences in the duration (four days versus one month) and scope (company channel only versus several public and semi-public channels) of the two collection efforts, the data size difference becomes quite understandable.

To get a rough idea of how successful the data collection efforts were in terms of representativeness, we examined the gender distributions of the avatars present in each data. Avatar gender was chosen as the diagnostic variable, because it is accessible without further analysis, and because gender has time and again been shown to be an important structural variable (Räsänen, 2008a; on the significance of gender roles in online communication, see e.g. Christofides, Islam and Desmarais, 2009). If the sampling is representative, the gender distribution of the avatars present in the data should correspond to the gender distribution of the avatars in the back-end data. If the distributions differ significantly, the sample is biased. Even if a sample is representative on the gender dimension, it can still very well be biased in some other significant way, so this must be considered only a rough heuristic test of the sampling.

Method	Back-end data		Listening posts		Data donors	
	Cases	Proportion	Cases	Proportion	Cases	Proportion
Male	74245	56.7 %	402	54.3 %	135	56.0 %
Female	56693	43.3 %	338	45.7 %	106	44.0 %
Total	130938	100.0 %	740	100.0 %	241	100.0 %

Table 1. Avatar gender distributions in data sets collected from the server Euros

Table 1 presents the results from the server Euros. The gender distribution of the listening post data is not significantly different from the gender distribution calculated from the back-end data (two-tailed one-proportion z-test: $z=-1.306$, $p=0.192$). The gender distribution of the donor data is likewise not significantly different from the gender distribution in the back-end data (two-tailed one-proportion z-test: $z=-0.215$, $p=0.830$).

Method	Back-end data		Listening posts	
	Cases	Proportion	Cases	Proportion
Male	55923	54.6 %	390	51.8 %
Female	46441	45.4 %	363	48.2 %
Total	102364	100.0 %	753	100.0 %

Table 2. Avatar gender distributions in data sets collected from the server Zephyros

Table 2 presents the results from the server Zephyros. Only listening post and back-end data are available, as data donors were not sought on this server. As on Euros, the gender distribution of the listening post data is not significantly different from the gender distribution calculated from the back-end data (two-tailed one-proportion z-test: $z=-1.565$, $p=0.118$). The data also allows us to compare avatar gender distributions between the two servers. Interestingly, the gender distributions on Euros and Zephyros as calculated from back-end data are statistically significantly different (two-tailed two-proportion z-test: $z=9.986$, $p<0.001$).

Conclusions and discussion

In this article, we introduced three approaches to collecting user-to-user communications data from virtual environments, considered their representativeness and reliability, and discussed the ethical issues related to them. We also described a case study where the approaches were used in practice, and compared the resulting data sets with each other. In this section, we discuss the results and draw some conclusions in the form of suggested practices.

For all but the most opportunistic or explorative research, it is probably best to start from the research problem and determine what kind of data is needed based on the research question. This step includes defining the statistical population: what is the larger group or groups to which conclusions from the sample are to be generalized? Based on this definition, an appropriate scope for the sampling can be determined. Dimensions that should be considered when deciding what to sample include things such as what game or virtual environment, which servers, which communication methods or channels, which time periods, and what times of the day.

The next step is to determine which of the three data collection approaches is used for the actual sampling. The results of the empirical study suggest that all three approaches are viable. The listening post and data donor approaches both yielded sizable data sets that are

large enough to facilitate further processing through coding and statistical analysis. Both approaches also yielded avatar gender distributions that were very similar to the back-end data. This suggests that both were able to achieve representative sampling, although the sampling could still be biased on some dimension other than avatar gender; to what extent this is a problem depends on the theory being tested, and what variables could plausibly affect the outcome.

As all three approaches seem statistically viable, the decision on which approach to use depends largely on the scope of the desired sample and on practical factors such as available resources and research timetable. Back-end communication records obtained from the operator of the environment obviously provide the widest scope of data, but getting access to them can be very difficult and time-consuming, and may result in contractual restrictions on what can be published. Researchers should also make sure that using such records is legal and ethical. The researcher may be subject to different standards as the operator. If necessary, the operator should be asked to prune, anonymize or encode data before making it available.

As an alternative to operator-provided records, researchers can collect their own data sets using the listening post approach. This gives full control over the data and sampling method to the researcher, but requires considerable effort and resources to carry out in a scale comparable to the empirical study presented above. Longitudinal data moreover takes time to collect. Depending on the client program or the protocol used by the platform, it may be possible to automate parts of the collection process. If collecting data from a pay-to-play MMO such as *World of Warcraft*, researchers have to consider how to cover the license and subscription fees.

In our empirical study, listening post avatars entered player-run companies to get a view of behavior inside these groups. The avatars acted in "disguise" insofar as they did not disclose their researcher identity. The ethicality of this practice is a complicated question. Groups in UWO are quite open to new members and the discussions thus relatively public. In some other games, such as *EVE Online*, players may consider their group discussions extremely sensitive, due to the intensely competitive nature of the game. On the other hand, infiltrators attempting to spy on these discussions are a normal and expected part of the gameplay. Should this have any bearing on researchers' acceptable practices in the game? In any case, it is unlikely that our clumsy listening post avatars would have been granted entry to *EVE Online* corporations in the first place. This can be seen as a limitation of this particular data collection approach, but it also means that it is difficult to inadvertently

commit serious privacy infringements using this approach.

The approach that probably involves the least effort on the part of the researchers is to request users to donate their communications records for research purposes. If the researchers are lucky, even sizable longitudinal data can be obtained in an instant. However, this approach carries the greatest risk of yielding bad data due to self-selection bias, social desirability, and awareness of being observed. To mitigate the possible impact of these risks on research results, it is a good practice to exclude the donors' own conversations from analysis. An exception must be made if the research calls for private conversations to be included in the sample. From the perspective of sampling representativeness, tapping into back-end records would be a better approach to sample private conversations. However, from a legal perspective, the data donors approach is probably less problematic. Laws that protect the privacy of correspondence prevent third parties from intercepting private messages, but they do not prevent the legal recipient from forwarding the messages to a third party, or making them public. Still, researchers should consider the privacy and dignity of conversation partners. Especially if fragments are published, they should be anonymized.

Finally, regardless of which data collection approach is chosen, the empirical study highlighted an important point to keep in mind regarding the sampling frame: it cannot be assumed that one server is necessarily representative of the whole MMO or virtual environment. The avatar gender distributions on the two servers examined in the study, as calculated from back-end data, were found to be significantly different. With such large sample sizes, the z-test will indicate almost any difference as being statistically significant, but the difference has practical significance as well, as its magnitude is more than two percentage points.

Why would avatar gender distributions differ between otherwise identical servers? One theory that can be put forward explains the difference through server age. Euros, the server with the greater proportion of male avatars (Table 1), was launched two years before Zephyros (Table 2). Studies of technology adoption indicate that the first adopters of a new technology or medium are often predominantly male, while the proportion of female adopters increases in time as technologies gain acceptance (Räsänen, 2008b). Most people use an avatar that corresponds with their physical gender (Roberts, 1999, Hussain and Griffiths, 2008). As a result, older servers should have more male avatars, while newly established servers should tend towards a flatter distribution. The conclusion from this is that servers should not be examined as if they were parallel copies of each other, each representative of the whole. Each server has its own history that can have an impact upon the data being

collected.

References

Blascovich, J., Loomis, J., Beall, A., Swinth, K., Hoyt, C., and Bailenson, J.N. (2002). Immersive virtual environment technology as a methodological tool for social psychology. *Psychological Inquiry,* 13, 103-124.

Bos, N., Karahalios, K., Musgrove-Chávez, M., Poole, E. S., Thomas, J. C., and Yardi, S. (2009). Research ethics in the facebook era: privacy, anonymity, and oversight. In *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems*, pp. 2767-2770, New York: ACM.

Bruckman, A. (2002). Studying the amateur artist: A perspective on disguising data collected inhuman subjects research on the Internet. *Ethics and Information Technology,* 4(3), 217-231.

Bruckman, A. (2006). Teaching Students to Study Online Communities Ethically. *Journal of Information Ethics* 15(2), 82-98.

Castronova, E. (2006). On the Research Value of Large Games: Natural Experiments in Norrath and Camelot. *Games and Culture,* 1(2), 163-186.

Castronova, E., Williams, D., Shen, C., Ratan, R., Xiong, L., Huang, Y., and Keegan, B. (2009). As real as real? Macroeconomic behavior in a large-scale virtual world. *New Media and Society,* 11(5), 685-707.

Cheseboro, J. W., and Bonsall, D. G. (1989). *Computer-mediated communication: Human relationships in a computerized world*. Tuscaloosa: University of Alabama Press.

Christofides, E., Islam, T., and Desmarais, S. (2009). Gender stereotyping over instant messenger: The effects of gender and context. *Computers in Human Behavior,* 25, 897–901.

Creswell, J. W. (2003) *Research design: Qualitative, quantitative and mixed methods approaches.* London: Sage.

Ducheneaut, N., Yee, N., Nickell, E., and Moore, R. J. (2006). "Alone together?": exploring the social dynamics of massively multiplayer online games. In: *Proceedings of the SIGCHI conference on Human Factors in computing systems (CHI 2006)*, pp. 407-416, New York: ACM.

Glaser, B. G. and Strauss, A. L. (2006). Theoretical Sampling. In N. K. Denzin (ed.), *Sociological Methods: A Sourcebook,* pp. 105-114, New Brunswick, NJ: Aldine Transaction.

Hair, J. F., Black, B., Babin, B., Anderson, R. E. and Tatham, R. L. (2006). *Multivariate*

Data Analysis (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Harris, H., Bailenson, J. N., Nielsen, A., and Yee, N. (2009). The Evolution of Social Behavior over Time in Second Life. *PRESENCE: Teleoperators and Virtual Environments,* 18(6) (forthcoming).

Hudson, J.M. and Bruckman, A. (2005). Using empirical data to reason about internet research ethics. In *Proceedings of the 2005 Ninth European Conference on Computer-Supported Cooperative Work (ECSCW)*, pp. 287-306, London: Springer.

Hussain, Z. and Griffiths, M. D. (2008). Gender Swapping and Socializing in Cyberspace: An Exploratory Study. *CyberPsychology and Behavior,* 11(1), 47-53.

Johnson, M. and Sihvonen, T. (2009) On the Dark Side? Gothic Play and Performance in Virtual Worlds. *Journal of Virtual Worlds Research,* 1(3). http://journals.tdl.org/jvwr/article/view/368

Kleinberg, J. M. (2007). Challenges in mining social network data: processes, privacy, and paradoxes. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 4-5, New York: ACM.

Kozinets, R. V. (2006). Netnography 2.0. In R. W. Belk, U. N. Cheltenham and M. A. Northampton (eds.), *Handbook of Qualitative Research Methods in Marketing*, pp. 129-142, Cheltenham: Edward Elgar Publishing.

Leuski, A. and Lavrenko, V. (2006). Tracking dragon-hunters with language models. *Proceedings of the 15th ACM international conference on Information and knowledge management*, pp. 698-707, New York: ACM.

McKee, H. A. and Porter, J. E. (2009). Playing a Good Game: Ethical Issues in Researching MMOGs and Virtual Worlds. *International Journal of Internet Research Ethics,* 2(1), 5-37. http://ijire.net/issue_2.1/mckee.pdf

Narayanan, A. and Shmatikov, V. (2008) Robust de-anonymization of large sparse datasets. In *Proceedings of 29th IEEE Symposium on Security and Privacy*, pp. 111-125, New York: IEEE Computer Society.

Neuendorf, K. A. (2002). *The Content Analysis Guidebook.* London: Sage.

Roberts, L. D. (1999). The Social Geography of Gender-Switching in Virtual Environments on the Internet. *Information, Communication and Society*, 2(4), 521-540.

Räsänen, P. (2008a). *In the Twilight of Social Structures.* Saarbrücken: VDM Verlag.

Räsänen, P. (2008b). The Aftermath of the ICT Revolution? Media and Communication Technology Preferences in Finland in 1999 and 2004. *New Media and Society,* 10(2), 225-245.

Shamoo, A. and Khin-Maung-Gyi, F. (2002). *Ethics of the use of human subjects in research: practical guide.* New York: Garland Science.

Shamoo, A. and Resnik, D. (2003). *Responsible Conduct of Research*. New York: Oxford University Press.

Steinkuehler, C., and Williams, D. (2006). Where Everybody Knows Your (Screen) Name: Online Games as "Third Places". *Journal of Computer-Mediated Communication,* 11(4). http://jcmc.indiana.edu/vol11/issue4/steinkuehler.html

Yee, N., Bailenson, J. N. (2008). A method for longitudinal behavioral data collection in Second Life. *PRESENCE: Teleoperators and Virtual Environments,* 17, 594-596.