Qualitative and Quantitative Methods in Libraries (QQML) 4:1001 –1018, 2014

Virtual Trace: A Framework for Applying Physical Trace Research Methodology in a Virtual Electronic Context

Frank Lambert, PhD [1]

[1] School of Library and Information Science, Kent State University, P.O. Box 5190, Kent, OH 44242, flamber1@kent.edu

Abstract: *Virtual trace* is presented as a reconceptualized methodological framework inspired particularly by Webb et al's physical trace methodology and a variety of Webometric data collection and analysis methods from the LIS literature. With the ongoing proliferation of data from current and new Internet-based sources, virtual trace is intended to be considered by experienced and novice researchers as a comprehensive research approach for studies whose designs are similar conceptually to those described in this article. Other online methodologies such as social tagging analysis and virtual ethnography are examined to provide virtual trace further definition.
Keywords: Virtual trace; Physical trace; Research methods; Web logs; World Wide Web; WWW

1. Introduction

The Internet, and the World Wide Web (WWW) particularly, offers both researchers and practicing professionals in the library and information sciences (LIS) an avenue to more than ample and valid data that may be used for a multitude of beneficial purposes. These purposes include applied and basic knowledge creation from a scholarly research perspective to more action-oriented research such as tools to facilitate decision making within information-based organizations and institutions such as libraries, museums, and archives. Lest we forget, other disciplines/professions may benefit equally from the exploitation of such data too. From business to communications to computer science and beyond, the data left behind by users of the wide variety of informational and transactional resources available on the WWW offer enormous benefits for knowledge creation of all types. While there is the potential that sinister motivations might exploit such data in ways that may leave individuals feeling quite insecure, the very good that may arise from

Received: 22.8.2013 / Accepted: 14.3.2014 ISSN 2241-1925

analyzing human behaviours in an online environment must be appreciated and promoted too.

A variety of questions may drive the research that exploit the data from a variety of Internet- and WWW-based data sources. A brief sampling of such questions might include: What type(s) of information are people searching for through a Web site, a search engine, or an online public access catalogue (OPAC)? What type(s) of information seeking behaviour(s) do these same people exhibit while searching for that information? Through the tagging of resources available on Web sites such as online images and other electronic media, what do people think that those entities mean? What are those entities "of" or "about?" These are but a few illustrative example questions that may arise from curious investigators. The answers to these questions based on the analysis of the data produced by human activities while engaged in online tasks will hopefully benefit in some way the ongoing development of online information retrieval systems, improve generally online information literacy, and increase knowledge overall in regard to information seeking behaviours, needs, and uses, information organization, and information systems.

2. Problem Statement

Since so much research in a wide variety of disciplines (e.g. Communications, Geography, LIS, Business, etc.) has turned to the Internet and the WWW for sources of new data to solve new and interesting action, applied, and basic research problems, it would be helpful for these particular researchers to have access to an appropriate methodological framework to help them in their studies. That is this paper's rationale. It presents a rearticulation of a current, interesting, and perhaps under-appreciated/used (based on its infrequent mention in the LIS literature) research methodology known as physical trace, but it does so within the framework of inquiry that is described above. This is done by reintroducing readers to physical trace methodology but *with a contemporary data source reorientation*. To do this, a comprehensive review of a variety of research/data collection methods used by researchers in the library and information sciences is presented to provide concrete examples that would support this rearticulation of physical trace into what is called now *virtual trace methodology*. The virtual trace referred to here is the evidence of traces of humans' activities stored in electronic form that has been created as a result of their actions in the virtual world defined as, and within the ephemeral boundary referred to, as the Internet generally but more specifically as the WWW. While the orientation of this paper is towards the LIS field particularly, this does not mean that this rearticulated (not "new") methodology is exclusive to this disciplinary domain. It just so happens that various data collection and analytical methods that may be considered to fall under the virtual trace methodological umbrella have been used with great success in LIS. Additionally, this paper will serve hopefully as a catalyst to the development of a rich collection of intellectual materials from other disciplines that will develop

and refine virtual trace further as a more effective methodological framework of inquiry.

3. Origins: What is Physical Trace Methodology?

Physical trace methodology was articulated and published first as a chapter in *Unobtrusive measures: Nonreactive research in the social sciences* by Webb et al in 1966. This monograph has been republished since, most recently in 1999 (Webb et al, 1999). During that 33 year span, the basic premises of physical trace as an empirical research methodology has not changed at all.

Webb et al noted how research studies up to the 1960s emphasized data collection methods that opened themselves to problems of "reactivity" or "centre of attention" effects, "because the respondent or participant in such research is typically aware of being observed" (Palys, 2003, p. 228). As a result, Webb et al suggested that researchers consider changing the scope of the data variables and their indicators to resolve their respective research questions. This would be done through the application of data collection methods that do not involve any sort of interaction with human research subjects such as variable manipulation typical in experimental design, surveying through some form of direct or indirect questioning, or observing human behaviours in natural settings through different levels of interaction and/or involvement. In other words, researchers should consider engaging in unobtrusive research as a way of studying social behaviour so that it does not impact, bias, or have an effect on the social behaviour under investigation (Babbie, 2011).

While non-reactive data collection measures sounds attractive because of potential bias minimization due to the absence of the researcher's presence, physical trace itself does not receive usually much space in research methods textbooks. It is presented often in a chapter on unobtrusive methods but only takes up roughly 3 or fewer pages of the text (e.g.; Berg, 2009; Babbie, 2011; Palys, 2003). Typically, physical trace is given less consideration than other unobtrusive methods such as content analysis or archival/historical methodologies.

It is possible that physical trace's scant review in research methods texts is due to potential problems with the methodology. Webb et al. (1999) acknowledge that, "it should be emphasized that physical evidence has greatest utility in consort with other methodological approaches." (p. 36) This means that physical trace data prove more valid when confirmed or triangulated through other observational data collected by the researcher(s) through different means, a method of strengthening the internal validity of the primary data collection method. However, this is true of other research methods too since the design of many social science research projects do not lend themselves well to external validity. As a result, all such studies may be strengthened usually by adopting a mixed-methods approach if it will address the respective research questions more completely. There also is then an implication that physical trace data are

not valid on their own, a problematic perspective in the forensic sciences. Related to this is that the actual chapter on the subject of physical trace is introduced by Webb et al as a methodology with a strong connection to physical evidence used in criminal investigations (Webb et al., 1999). It is interesting that Webb et al. never link explicitly their methodology to theory such as Locard's principle which states that "every contact leaves a trace." (Aitken and Taroni, 1995; p. 13) Locard's principle forms the basic core of forensic science, and perhaps Webb et al. were concerned that physical trace would be associated too closely with forensics and thus wished not to overemphasize this association. Regardless, Webb et al. were thinking more empirically with their text and their respective physical trace chapter especially in terms of solving larger research problems rather than determining one's guilt or innocence. The potential weaknesses of physical trace methodology need not be transferred necessarily to virtual trace thanks to other data attributes that are simply not available with physical trace data. And, Locard's principle need not be limited to physical trace either. In fact, virtual trace may facilitate the exercise of Locard's principle.

Physical trace methodology is comprised of two research methods that are not necessarily mutually exclusive; the study of the accretion and/or the erosion of physical traces of human behaviour. While the former is more appropriate in the context of virtual trace, both methods should be discussed.

Erosion measures might be considered the most obvious type of physical trace, and yet is arguably the least likely to be applied in virtual trace. Usually, most people's association with erosion has to do with the slow and deliberate destruction of natural physical entities such as earth, rocks, etc., caused by wind and/or water. This same principle applies in erosion measures in physical trace. It is essentially the wearing away of all sorts of different types of physical objects. However, rather than being caused by natural elements, this erosion is the result of human activities, decisions, and/or other social/economic factors that might affect those same persons. Webb et al. (1999) use a couple of examples to demonstrate erosion: the wearing away of floor tiles in front of popular museum exhibits; or, the wear on library books that might indicate more concretely what is a "popular" book instead of relying simply on one data source such as circulation data. However, natural erosion does have a role of sorts in providing social and economic data that may be of use to researchers. For example, consider the effect of natural erosion has on 19th-early 20th century grave markers as a data source. Typically, more financially fortunate persons were able to afford stone grave markers upon their death whereas less financially fortunate persons were able to afford only wood grave markers. Thus, the proportion of wood to stone grave markers gives the researcher first a snapshot of the historic economic demographics of a community. However, since the wood markers would erode and degrade much quicker than the stone markers, researchers could not rely solely on grave markers as an indicator of the same community's health and life expectancy because the stone markers would provide data only of the more financially fortunate persons. Since those

persons could likely afford better health care, their life expectancy could be longer than that of those less financially fortunate. Thus, natural erosion as applied to physical trace may indeed provide a biased snapshot of this example community in terms of a life expectancy variable while at the same time providing an accurate snapshot of the community's economic demographics.
While erosion is the wearing away of objects based on human activity especially, accretion measures are the accumulation of traces of human behaviours. For example, while the wearing (erosion) of tiles in front of a museum display may be indicative of its popularity, the accumulation of fingerprints on the glass in front of the display (accretion) is just as effective an indicator. In some ways, erosion and accretion work together for analytical purposes. In fact, the fingerprint example could give a more detailed indication of who is looking at the display, with the lower and likely smaller fingerprints indicating children, and the higher and larger ones adults (assuming adults are as likely to place their hands on the glass as children). Additionally, in libraries, the accumulation of dust on library materials may indicate a lack of popularity for the respective item (Webb et al., 1999). However, the accretion of dirt from hands on books' pages is equally indicative of popularity as is the wear on said books. Also, the accumulation of books on tables in the libraries is another way of looking at the "use" of materials in a library beyond circulation records.

Like *any* other research method, physical trace has disadvantages. One such disadvantage is in the realm of research ethics, but that is dealt with in a later section along with possible solutions. The other disadvantages have to do with selective deposit and selective survival, topics that have applicability to other methods such as archival research. Selective deposit has to do with the abilities of certain social groups, within certain societies, to preserve the materials related to those groups' activities based on available resources, power, education, etc. For example, millennia from now, which record and perspective of the relationship between Native North Americans, a minority group whom has a strong oral history tradition, and North American governments, with a strong records management/archival tradition, will survive? This has nothing to do with the veracity or accuracy of the perspective of one side or the other. This comes down to issues of power and money, both of which will have a significant effect on whose position is heard best (Palys, 2003). Selective survival is based on decisions, resources, and (economic) power too, as exemplified in the grave marker example above. For instance, clay lasts longer than wood; alkaline paper does not degrade as quickly as acid paper (Webb et al, 1999). Thus the decisions and other variables that affect selective deposit and selective survival could cause unknown unrepresentativeness and bias in the future when these traces are examined for a research study.

Both principles of erosion and accretion in physical trace go beyond examining merely books in libraries, and beyond library and information science in general. In the spirit of this paper's topic of virtual trace, the machines that create the virtual trace also are just as subject to physical trace measures of

erosion and accretion; except, this is rather limited to whether computers are being used in a library or not. Buildup of dirt on keyboard keys, and the wearing of the imprinted letters on said keys, *is* evidence of use. The important question is not answered though: for what are those computers being used? This is where virtual trace and its associated data collection methods are relevant.

4. Physical and Virtual Trace in Context

One advantage virtual trace possesses over its "cousin" physical trace is that it has additional attributes of digital data to offer context to the units of analysis researchers may focus on. In physical trace, a garbage receptacle may be referred to arguably as a technology as some sort, in much the same way as Robert Moses's bridges in Long Island (Winner, 1999). Radios may be considered to be more "typical" of a "technology" in that it requires some form of electric charge in which to operate (cf. Webb et al's (1999) reference to the radio-dial study that estimates radio station popularity by sampling cars in automotive service departments). However, in the case of a garbage can or a radio as a source of data for physical trace, neither offers additional data points that may provide important contextual data that one may find often in virtual trace data. For instance, if a researcher is examining physical trace measures of erosion and accretion to define "use" of library resources by visitors, the only conclusion that the researcher may make in some cases is that various parts of the library are being used. Wear patterns in the floor shows how traffic moves in the library, maybe indicating popularity of resources, but never confirming on its own that this is indeed the case. Dust buildup on certain physical resources (e.g., books or videos) may indicate lack of popularity or desire for particular specific portions of the library's collection, but again it never confirms on its own that this is indeed the case. There are exceptions to these weaknesses in physical trace methods, as the fingerprint example mentioned earlier, but that is because there is the possibility of other data being in the vicinity. However, while physical trace offers many important clues that can help lead to new, or even to refine, a whole variety of research questions, physical trace is just as likely to also leave serious gaps in the validity of researchers' conclusions without examining the trace data in context with the findings of different data with the same goals that are collected through other methods.

This is where virtual trace data may offer a substantial advantage over physical trace data. Virtual trace makes it relatively simple, while being unobtrusive still, to collect other data points for other variables while collecting the primary data that are needed to address the original research problem and its associated research questions/hypotheses. Compare this to physical trace using specifically the example of wear patterns in the floor of a library. While this erosion measure indicates perhaps the "use" or "popularity" of library resources, it does not tell the researcher WHO created this wear (e.g., gender, age, ethnicity/national origin, education, annual income, etc.), WHEN the wear occurred, or WHY the wear created without knowing additional context (e.g.;

had a display changed, or had the collection been reorganized physically?). Some additional variables that may potentially be included in virtual trace data include WHEN the person was engaged in some activity while online with a particular Internet or WWW application, WHAT the person was doing, HOW the person was doing their particular activity, WHY, through reasoned implication often, the person was doing what he/she was doing, WHERE the person was located physically when engaging in this activity, and, very rarely due to the anonymous nature of the WWW, WHO this person is. While it is rare to find all of these data attributes being recorded in all cases, there certainly is the potential to configure data collection tools to collect all of these variables.

Besides the methodological limitations mentioned earlier in relation to selective deposit and selective survival, physical trace methodology also has empirical limitations; it will not tell the full story of the particular phenomena that the researcher is seeking to understand. The same is true too of virtual trace. As with physical trace, virtual trace may benefit through data triangulation that other data collection methods provide.

5. Demonstrable Virtual Trace Methods

Since this paper claims that physical trace methodology should be conceptualized in a new light for electronic trace evidence, it seems appropriate that secondary evidence of published research be presented that demonstrates methods that are exemplars of what is determined to be virtual trace methodology. The sample of studies presented below are published primarily in library and information science source media that use a variety of methods that reflect best virtual trace methodology and reflect especially the trace concept of accretion.

5.1. Navigational and Query Web Log Analysis

Certain virtual trace data collection methods such as Web log analysis might be described as an Webometric data collection method. In fact, Thelwall, Vaughan, and Björneborn (2005) imply that Web log analysis, or also referred to as Web log data mining, might lie within the Webometric domain. This is because "in addition to studying the Web itself, it can also be useful to collect data on how humans interact with it: on how it is used. Data collected could include behavioral studies of humans using the Web" (p. 95) On its face, it does make sense that the analysis of attributes related to querying and searching on the WWW might "fit" within the methodological framework of Webometrics; for instance, George Zipf's work concerning the distribution of words in a text is cited often as a standard mathematical model along with the models of other bibliometricians (e.g., Lotka, Bradford) in classical bibliometrics (De Bellis, 2009). Since Zipf's distribution model is dependent on word frequencies, much like the analyses of queries submitted via search engines are dependent often on query term frequencies, it seems logical then that Web log analysis might see itself within the respective methodological umbrellas of Webometrics more narrowly and informetrics more broadly.

However, there is a rather distinct difference between Webometrics and a virtual trace method like Web log analysis. Webometrics more like classical citation analysis and bibliometrics particularly, is concerned with the tracing of links to Web sites of a particular type from other Web sites. As Thelwall and Vaughan (2004) write, "Webometrics encompasses all quantitative studies of Web-related phenomena. … Hyperlink studies typically draw on the more mature field of bibliometrics for ideas, providing a connection to information science" (p. 1213). Also, there is a certain level of intentionality related to the act of creating links from one Web site to other Web sites. What may not be clear from the analysis of these links is why the Web site designer made these links. There is a certain degree of inference that must be made to determine why these links were made whereas meaning that Webometric analysis may not be ideal method for determining online behaviour beyond "Web site X linked to Web site Y because they both deal with Z."

The reliance on frequencies to aid in analysis of virtual traces of human activities on Web sites is evident in both navigational Web log studies and research of Web logs that capture information seekers' queries expressed through the submission of terms. These studies also use multiple attributes of the Web logs housed on servers to provide much more detail about search sessions in the aggregate that do not exist necessarily with "pure" Webometric or with physical trace studies. Many of these attributes capture data without the information seekers' knowledge too, further reducing potential bias in the overall analysis of Web logs beyond the queries themselves. It should be remembered though that like with any other research method, Web log query analysis does have its limitations (e.g. Spink and Jansen, 2004; Thelwall, Vaughan, and Björneborn, 2005). However, it also has serious advantages. For example, it "can reveal first-hand and real-world behavior and interests of [Web site] users. It enables researchers to better understand Web site user behaviors and the service quality that the Web site provides. It also can be used to optimize the effectiveness of information services." (Zhang et al., 2008, p. 1934) The observation of first-hand and real world behaviour and interests of users in Web log data is possible due to the unobtrusive nature of the data collection using this method (Spink and Jansen, 2004). The primary attribute of unobtrusiveness links Web log analysis methods strongly to the established unobtrusive methodology of physical trace. However, considering the primary units of analysis in Web log analysis are far from physical, and potentially more ephemeral than many traces in our physical world, it seems appropriate to articulate a new methodological perspective; hence, virtual trace.

Two research streams have emerged in Web log analysis. One is the examination of users' navigation behaviour within a Web site as they move from one page to another (e.g., Bertot, McClure, Moen, and Rubin, 1997; Nicholas, Huntington, Lievesley, and Wasti, 2000; Nicholas, Huntington, Williams, Lievesley, Dobrowolski, and Withey, 1999; Nicholas, Huntington, Lievesley, and Withey, 1999; Nicholas and Huntington, 2003). In other words,

the clicks from a person's mouse pointer on links within Web sites are recorded and logged. These virtual traces of links visited within a Web site and recorded duly by software thus represents a person's need to know or interest in what the Web site's respective link represents as a topic and are thus the primary data points of interest. However, the data consisting of links that information seekers have visited including their own final information destination may be fraught with misleading indicators. This is because in part a point-and-click information access paradigm as an information retrieval practice is dependent very much on the starting page and the information the seekers need to continue their journey in the Web site (Herrera-Viedma & Pasi, 2006). It is up to the Web site designer to name the site's pages and build content. Additionally, the terminology and the information content contained on the final destination page may make a determination of what type of information the seeker was seeking difficult.

The other research stream is the scrutiny of search engine log files where data are recorded about what users are actually seeking as represented by the actual keyword queries that users submit (e.g., Thelwall, Vaughan, and Björneborn, 2005; Silverstein et al., 1999; Spink et al., 2001; Beitzel et al., 2007; Chau, Fang, and Sheng, 2005; Ross and Wolfram, 2000; Wang, Berry, and Yang, 2003; Pu, Chuang, and Yang, 2002; Lau and Goh, 2006; Rieh and Xie, 2006). As Spink and Jansen (2004) elaborate helpfully, in Web log analysis "[t]erms are the basic building blocks through which a Web searcher expresses their

Moen, and Rubin (1997) applied many of these techniques in a study of a medium-sized federal agency's Government Information Locator Service. While also reporting the results of their study, the authors also defined four logs of particular importance that are created by log analysis software: access, agent, error, and referrer logs. In addition, they also make the noteworthy distinction between mere hits or files downloaded from a Web site by the user, and accesses, or page impressions, which is the download of a whole page, since early Web log analysis software would only record the former variable's values.

Nicholas, Huntington, Lievesley, and Wasti's (2000) yearlong case study using a total of four months of log data from *The Times/Sunday Times* newspaper also adopted many of the approaches published by Bertot, McClure, Moen, and Rubin (also detailed in Nicholas, Huntington, Williams, Lievesley, Dobrowolski, and Withey, 1999; Nicholas, Huntington, Lievesley, and Withey, 1999). However, what distinguished this study from others was that the authors had access to independent variables such as the respective Web site user's gender, country of origin, and type of organization. This came from the personal information that the user submitted when he/she originally subscribed to *The Times* newspaper and Web site. Having to log in prior to using the Web site, individual "identification" from these basic values could be used by the researchers. Since some characteristics were already known about the users, they were not completely anonymous as would be the case in a study of the use of a Web site without some form of log in. However, research ethics undoubtedly meant that the authors could not publish any identifiable data about the users of *The Times* Web site.

Nicholas and Huntington (2003) also developed different approaches to log file analysis; specifically, the demonstration of the use and utility of 'micro' techniques that examine sub-groups to improve the analysis of Web statistics in order to address the problems associated with resolving IP numbers to a single user and cross-border IP address registration. The intent was to derive as much detailed information "from the *skeletal digital fingerprint that people leave behind* (emphasis added) when they access the service" (Nicholas and Huntington, 2003, p. 403). Using SPSS to analyze the log files of a health Web site, the authors scanned sub-groupings to find particular user groups. A reverse domain name server (DNS) lookup on the incoming IP numbers was used to find the user-type information since the DNS can inform about the geographical registration of the user as well as the type of user (Nicholas and Huntington, 2003). Besides examining the overall geographic distribution of incoming requests, they also analyzed the level of activity during a session, the number of visits by return users, and tracked four randomly selected return users individually by IP address and client browser as individual cases.

5.3. *Query Term Web Logs*

Research published on the analysis of Web query logs has focused on querying behaviours through either commercial search engines, search engines that are

used exclusively for searching through a Web site, or for searching through library OPACs. Silverstein et al. (1999) examined a data set recorded over 43 days from the AltaVista search engine comprising over 900 million total requests. Their interest lay in determining which queries were most common, the average length of queries, how many queries were submitted during an individual session and, especially, correlations between query terms and other field values. Spink et al. (2001) conducted a similar study involving over one million query logs (531,416 of which were unique) sent to the Excite search engine on one day by 211,063 users. They examined the number of queries submitted per identified user, measured the change in unique queries submitted by each identified user, the number of results pages viewed, and whether multi-term queries used advanced search features such as Boolean operators. Rieh and Xie (2006) also analyzed a data set of 313 search sessions from the Excite search engine to characterize facets of query reformulation and identify patterns of multiple query reformulation in Web searching. Their goal was to explore "the ways in which search engines can support query reformulation more effectively in Web searching" (p. 752) by using Saracevic's stratified interaction model as an analytical framework (Rieh and Xie, 2006).

Ross and Wolfram (2000) used the frequency of binary term co-occurrence following the parsing of multi-term queries to determine facets of multi-term queries without having to look at every query. This was a bottom-up grounded theory method with no *a priori* categorization beforehand. Pu, Chuang, and Yang (2002) used this same approach to develop their subject taxonomy for the automatic classification of Web query terms into broad subject categories. Beitzel et al.'s study (2007) involved the analysis of the Web query logs of the America Online (AOL) Web search service. However, the authors used a longitudinal analysis "to examine static and topical changes [in querying] over longer periods such as days, weeks, and months" (Beitzel et al., 2007, p. 167). Additionally, the authors "analyzed the queries representing different topics using a topical categorization of [their] query stream" in an effort to determine how querying behaviour for some categories would either change or remain static over time (p. 167). Lambert (2010a, 2010b, 2012) used methodological approaches similar to Ross and Wolfram, Pu, Chuang, and Yang, and Beitzel et al. to examine information seeking needs and behaviours expressed through querying on community information and municipal government Web sites. However, Lambert's studies (2010b, 2012) examined such information seeking over a period of three years, one of the longest time periods in the literature.

Chau, Fang, and Sheng's study (2005) focused on keyword queries that were submitted through the Utah state government Web site. The authors' goal was not only to determine the characteristics of queries that were submitted through this Web site search engine but also to compare those same queries to those submitted through general-purpose search engines. They determined "that Web users behave similarly when using a Web site engine and a general-purpose search engines in terms of the average number of terms per query and the average number of result pages viewed per sessions*(sic)*" (p. 1374). However,

the users of the Utah state government Web site search engine submitted, on average, fewer queries per session than users of general-purpose search engines. Additionally, Utah government Web site users use different sets of terms and topics in their queries compared to general-purpose search engine users.

Lau and Goh (2006) analyzed 641,991 queries from the Nanyang Technological University OPAC to determine what caused failed search sessions. Their objective was to identify areas of improvement for the OPAC to improve users' search experiences with the OPAC through the use primarily of failure analysis. As a result, the authors recommend improvements to the OPAC through enhancements to interactive query reformulation, browsing, and context-sensitive assistance.

Highlighted in the studies above is that Web log analysis, regardless if it relies on data captured of Web site users navigating through links or if general users of the WWW submit queries expressed through key words to help them find the information they seek, is a flexible and evolving approach to determining how persons search for information and the type(s) of information sought. What needs to be emphasized in these same studies is that the information seekers'/Web site users' various contacts with the respective Web-based media have left an electronic, virtual trace. Additionally, these accumulated (accretion) virtual traces are being left behind with likely very little knowledge by the information seekers/Web site users that they are doing so in the first place. This is much like what happens with contributors to physical trace. It is these realizations of rather innocuous behaviours such as setting one's radio station or disposing of refuse that lead to nearly "pure" or nearly unbiased data that demonstrate so much about human behaviours. However, this need not be limited only to the tactile, physical world through which humans have displayed a myriad of behaviours. Considering how important the electronic, virtual world has become for so many contemporary entities, it seems appropriate that a modified framework such as virtual trace be articulated and built upon further to continue the research of the online behaviours of WWW and Internet users.

An important attribute of the virtual traces left behind in Web logs hosted on organizations' servers has to do with *intentionality*. The modified paradigm of virtual trace adopts this attribute of physical trace in that the traces resulting from human behaviours are, in almost every circumstance, *unintentional* or even unconscious. These result from actions that have been conditioned based on a myriad of experiences, social norms and pressures, etc. This attribute may not happen in every instance. However, intentionality is an important facet of virtual trace in that it distinguishes it from other potential methodologies that would be effective in empirical observation of other online behaviours.

6. Related Virtual Trace Approaches: "Virtual" Participant Observation; and, Social Tagging and Analyzing Folksonomies

Studies of online behaviour need not rely solely on Web log analysis. Virtual trace, as defined and demonstrated above, relies primarily on the electronic traces left behind of those whom are actually searching or seeking for information. However, other methods, such as ethnography or content analysis, also have been used to analyze other electronic traces left behind. The big difference between Web long analysis and these new virtual methods has to do with the Web users' intent and expectation of what they "produce" virtually. The virtual traces of online behaviours such as keyword querying are not visible immediately to, nor accessible immediately for, other Web users to view; the recording of the behaviours go on "behind the scenes." However, there are many, many instances of online behaviour where this is NOT the case. Through posting comments on blogs, making contributions to wikis, or tagging Web pages with keywords to describe what they think the page represents or is about, many users of the World Wide Web and its myriad resources leave their own virtual traces behind *intentionally* for those whom may find them of interest to view and interact with consequently. This interaction with others on the Web through Web-based resources is the premise behind so-called "Web 2.0" resources such as blogs, wikis, and social tagging activities. "Web 2.0" has allowed regular WWW users to make public various materials on the WWW without having to create their own Web sites or Web pages.

As a result of the apparent difference in the type of data found in Web logs, it may be prudent to limit initially the paradigm in which virtual trace may be defined because studies of other types of online behaviours, such as blogging, posting to wikis, tweeting, tagging, etc., may be studied reasonably successfully within other methodological frameworks such as observational ethnography or other unobtrusive methods such as content analysis. As with virtual trace, the published literature that alludes at least to these other frameworks needs further thought and development to provide future researchers clearer methodological structures for their respective studies.

Sade-Beck (2004) articulated a qualitative framework for ethnographic research conducted via the Internet and the WWW specifically. This framework was developed through an empirical study focusing on Israeli support communities on the WWW. Rather importantly, Sade-Beck writes nearly immediately that this "virtual" ethnography should be implemented through multiple data collection methods; in this case, online observations, interviews, and the content analysis of materials germane to Sade-Beck's particular research problem. Several advantages are highlighted with virtual ethnography, some of which mirror those found in virtual trace. Not surprisingly, virtual ethnography differs from face-to-face communication in that many of the physical attributes of communication are absent. However, the anonymity inherent in virtual communications through the use of pseudonyms in blogs, for instance, allows for more honest expressions of feelings, thoughts, and opinions, regardless of how personal they may be.

It is clear that virtual ethnography eliminates a major source of data; the physical communication persons engage in as they interact with one another either in pairs or in groups. Not only is the use of additional data collection methods important in "standard" ethnographic observation studies, but it is more important in virtual studies because the physical aspect is removed. Therefore, one may wonder whether virtual ethnography is indeed ethnography at all or if it is merely content analysis of virtual documents (e.g., blog posts, content on static Web pages, etc.) one may find on the WWW. Regardless, online ethnography is not necessarily a form of virtual trace research because intentionality becomes the primary differentiating attribute of the two respective methodologies. Research subjects such as those studied by Sade-Beck are leaving virtual traces intentionally to contribute to the Holocaust support communities. There also is a different written contribution to these online support communities not quite like that found in Web search query logs where one or two terms are submitted to find other information.

Intentionality becomes a characteristic in social tagging of Web pages too.

> "Social tagging, which is also known as collaborative tagging, social classification, and social indexing, allows ordinary users to assign keywords, or tags, to items. Typically these items are Web-based resources and the tags become immediately available for others to see and use. Unlike traditional classification, social tagging keywords are *typically freely chosen* (emphasis added) instead of using a controlled vocabulary. Social tagging is of interest to researchers because it is possible that with a sufficiently large number of tags, useful folksonomies will emerge that can either augment or even replace traditional ontologies." (Tonkin et al, Jan. 30, 2008)

The end result is that these folksonomies serve as a pseudo-controlled vocabulary that may assist in the indexing and retrieval of Web-based documents. Researchers use a number of methods to make some determination as to what characterizes various tagging behaviours. The consequent benefit of such studies is that better understandings of manual indexing through the use of formal controlled vocabularies may emerge, thus hopefully improving manual indexing behaviours in general. Researchers use a variety of metrics such as the number of tags given, tag co-occurrence, and measured frequency as methods for analysis (Tonkin et al, Jan. 30, 2008). In fact, many of the analytical methods are similar exactly to those applied in informetrics generally.

Like virtual ethnography, online social tagging may be considered in some ways to be a form of virtual trace. Taggers' virtual contacts with Web sites do leave virtual traces of their online presence and interactions that, as with blog posters whom are studied in virtual ethnography, are not known necessarily by the respective Web site visitor; and, these traces are recorded in various Web logs, recording how the visitor navigated through the Web site, how the visitor got to the Web site through a referring Universal Resource Locator (URL), and if the visitor queried the Web site to find directly pertinent information. All of these

actions occur behind the scene and quite unintentionally from the perspective of the WWW user. However, as with virtual ethnography, the intentions of persons whom visit Web sites to tag materials is clear. They are there to engage in a task that they know will leave a trace, that being tags. And that is what differentiates virtual trace and its respective analytical methods from virtual ethnography and social tagging.

7. Limitations and Ethics of Virtual Trace Research Methodology

Based on the examples of the studies cited above in order to offer a foundational research literature for virtual trace, it is possible to state that virtual trace is at its core a quantitative approach to the study of online information behaviours. For instance, calculating the frequencies of query term submissions, or calculating the sum of links clicked on by information seekers, over the course of one day, one week, one month, or one year, form the base for further analytical approaches to understanding a variety of information seeking behaviours. As a result of this analysis, researchers are confronted with potentially millions of queries or URLs. And, to be able to discern empirically generalizable types of online information behaviours requires researchers to consider these data points *en masse* rather than individually. Sampling thus becomes an important component of this data analysis.

One of the issues of analyzing virtual trace query log data is trying to identify accurately individual users from the search session identifiers. This is to ensure that duplicate query or navigational Web logs are not included for analysis which may result in data skew. For session identification, Web site designers have implemented the use of cookies because the proliferation of dynamic IP addresses has meant that it is impossible to identify individual users of a Web site (Rubin, 2001). While this is more accurate than relying solely on IP addresses for identifying these users, Web users may disable cookies on their computer, thus leaving this identifying field in Web logs blank and consequently affect individual query submission analysis (Thelwall, Vaughan, and Björneborn, 2005; Silverstein, Henzinger, Marais, and Moricz, 1999). However, considering that these log analyses involve hundreds of thousands or millions of queries, individual query submission analysis is a fool's pursuit *unless* that is what the respective research problem calls for. It is the aggregation of the hundreds of thousands-plus log entries along with virtually untraceable (to individual persons that is) IP addresses that provide the ethical assurances of anonymity and confidentiality when it comes to virtual trace. Include the fact that there is no interaction between the researcher and the anonymous research participants whom provide Web queries or navigational logs, and a nearly unassailable ethical research paradigm may be established. This is not dissimilar to researchers rummaging through garbage to find physical trace data. There is always the potential of identifying individuals, but the trouble involved likely serves no immediate research benefit. And, if other data are needed to add further empirical context, then other research methods always

may be used following a review of the research protocol by an institution's research ethics board.

8. Conclusions

Virtual trace methodology will serve hopefully as a helpful framework for current and future, and experienced and novice, researchers. As this paper has attempted to present, there are distinct data analysis methods to be associated with virtual trace, especially Web log analysis research conducted in the library and information sciences. What is needed to build upon this framework now is additional exemplar research from other disciplines that rely on the unobtrusiveness of this type of online methodology and its current, associated data analysis method. There is no shortage of digital, electronic, and *virtual* data currently that may be exploited ethically by researchers to help create new knowledge. There is likely to be new and unique such data in the future that result from the development and implementation of new online tools and applications. For instance, the social communication media Twitter shows promise as a data source for a variety of research problems. While the text contained in the 'Tweets' of Twitter users might be examined through a content analysis method, additional data attributes such as time and date, geographical location, etc., are likely to be recorded without the users' knowledge and thus verifying again that physical trace can serve as a useful methodology to present a complete picture of Twitter users.

Researchers in other disciplines will hopefully consider their own sources of digital data along with virtual trace as a methodological framework to add more to this discussion. All of the various methodological frameworks and their associated data collection methods used by social science researchers did not come from a vacuum. If virtual trace is to become a viable, useful, and rigourous way of creating new knowledge, then the contributions of many other academics from other disciplines to this hopefully ongoing discussion will be more than welcomed. With the contribution of other interested researchers, virtual trace will become a relatively cohesive, comprehensive, and well-defined methodological tool as researchers turn accordingly to the WWW and the Internet in their pursuit of knowledge.

References

Aitken, C.G.G. and Taroni, F. (1995). *Statistics and the evaluation of evidence for forensic scientists*. 2nd ed. New York: John Wiley and Sons.

Babbie, E. (2011). *The basics of social research* (5th ed.). Belmont, CA: Wadsworth.

Beitzel, S.M., Jensen, E.C., Chowdhury, A., Frieder, O., & Grossman, D. Temporal analysis of a very large topically categorized Web query log. *Journal of the American Society for Information Science and Technology*, 58(2), 166-178.

Berg, B. (2009). *Qualitative research methods for the social sciences* (7th ed.). Boston: Allyn & Bacon.

Bertot, J.C. & McClure, C.R. (1996). *Sailor Network assessment final report: Findings and future Sailor Network development*. Maryland State Department of

Education, Division of Library Development and Services. Accessed March 3, 2005, from http://research.umbc.edu/~bertot/sailor.final.report.pdf.

Bertot, J.C., McClure, C.R., Moen, W.E. & Rubin, J. (1997). Web usage statistics: Measurement issues and analytical techniques. *Government Information Quarterly*, 14(4), 373-395.

Chau, M., Fang, X., & Sheng, O. (2005). Analysis of the query logs of a web site search engine. Journal of the American Society for Information Science and Technology, 56(13), 1363-1376.

De Bellis, N. (2009). *Bibliometrics and citation analysis: From the Science Citation Index to cybermetrics*. Lanham, MD.: Scarecrow Press.

Herrera-Viedma, E., & Pasi, G. (2006). Soft approaches to information retrieval and information access on the Web: An introduction to the special topic section. Journal of the American Society for Information Science and Technology, 57(4), 511–514.

Lambert, F. (2010a). Online community information: The queries of three communities in southwestern Ontario. *Information Processing & Management*. 46(3), 343-361.

Lambert, F. (2010b). Web searching to meet everyday information needs: A comparative longitudinal study of queries submitted to an online community information system. *Prato Community Informatics Research Network (CIRN) Conference 2010: Tales of the Unexpected: Vision and Reality in Community Informatics*. CIRN -DIAC Conference: Monash University Centre, Prato, Italy 27-29 October 2010. Larry Stillman and Ricardo Gomez, eds.

Lambert, F. (2012. In press). Seeking information from government resources: A comparative analysis of two urban communities' Web searching of municipal government Web sites. *Government Information Quarterly*.

Lau, E.P. and Goh, D.H.-L. (2006). In search of query patterns: A case study of a university OPAC. *Information Processing & Management* 42(5), 1316–1329.

Nicholas, D., Huntington, P., Williams, P., Lievesley, N., Dobrowolski, T. & Withey, R. (1999). Developing and testing methods to determine the use of web sites: Case study newspapers. *Aslib Proceedings*, 51(5), 144-154.

Nicholas, D., Huntington, P., Williams, P., Lievesley, N. & Withey, R. (1999). Cracking the code: Web log analysis. *Online & CD-ROM Review*, 23(5), 263-269.

Nicholas, D., Huntington, P., Williams, P., Lievesley, N., & Wasti, A. (2000). Evaluating consumer website logs: A case study of The Times/The Sunday Times website. *Journal of Information Science*, 26(6), 399-411.

Nicholas, D. & Huntington, P. (2003). Micro-mining and segmented log file analysis: A method for enriching the data yield from Internet log files. *Journal of Information Science*, 25(5), 391-404.

Palys, T. (2003). *Research decisions: Quantitative and qualitative decisions* (3rd ed.). Scarborough, ON: Thomson Nelson.

Pu, H., Chuang, S. & Yang, C. (2002). Subject categorization of query terms for exploring Web users' search interests. *Journal of the American Society for Information Science and Technology*, 53(8), 617-630.

Ross, N.C.M. & Wolfram, D. (2000). End user searching on the Internet: An analysis of term pair topics submitted to the Excite search engine. *Journal of the American Society for Information Science*, 51(10), 949-958.

Rubin, J.H. (2001). Introduction to log analysis techniques: Methods for evaluating networked services. In C. R. McClure & J. C. Bertot (Eds.), *Evaluating networked information services: Techniques, policy, and issues* (pp. 197-212). Medford, NJ: Information Today.

Sade-Beck, J. (2004). Internet ethnography: Online and offline. *International Journal of Qualitative Methods*, 3(2), retrieved March 3, 2012 from http://www.ualberta.ca/~iiqm/backissues/3_2/ pdf/sadebeck.pdf

Silverstein, C., Henzinger, M., Marais, H. & Moricz, M. (1999). Analysis of a very large Web search engine query log. *SIGIR Forum*, 33(1), 6-12.

Spink, A., Wolfram, D., Jansen, M.B.J. & Saracevic, T. (2001). Searching the Web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3), 226-234.

Spink, A. and Jansen, B.J. (2004). *Web search: Public searching of the Web*. Dordrecht: Kluwer Academic Publishers.

Thelwall, M. and Vaughan, L. (2004). Webometrics: An introduction to the special issue. *Journal of the American Society for Information Science and Technology*, 55(14), 1213-1215.

Thelwall, M., Vaughan, L. and Björneborn, L. (2005). Webometrics. In B. Cronin (Ed.) *Annual Review of Information Science and Technology*, 39 (pp. 81-135). Medford, NJ: Information Today.

Tonkin, E., Corrado, E., Moulaison, H.L., Kipp, M.E.I., Resmini, A., Pfeiffer, H.D., Zhang, Q. (Jan. 30, 2008), Collaborative and social tagging networks. *Ariadne*. Accessed April 26th, 2012, from http://www.ariadne.ac.uk/print/issue54/tonkin-et-al.

Wang, P, Berry, M.W. & Yang, Y. (2003). Mining longitudinal Web queries: Trends and patterns. *Journal of the American Society for Information Science and Technology*, 54(8), 743-758.

Webb, E.J., Campbell, D.T., Schwartz, R.D., & Sechrest, L. (1966) *Unobtrusive measures: Nonreactive research in the social sciences*. Chicago: Rand McNally College Publishing Co.

Webb, E.J., Campbell, D.T., Schwartz, R.D., & Sechrest, L. (1999) *Unobtrusive measures*. Rev. ed. Thousand Oaks: Sage Publications, Inc.

Winner, L. (1999). Do artifacts have politics? In D. MacKenzie & J. Wajcman (Eds.), *The social shaping of technology, 2nd ed.* (28-40). Berkshire, U.K.: Open University Press.

Zhang, J., Wolfram, D., Wang, P., Hong, Y., and Gillis, R. (2008). Visualization of health-subject analysis based on query term co-occurrences. *Journal of the American Society of Information Science*, 59(12), 1933-1947.