

**MEASURING ONLINE JOURNAL USAGE:  
ISSUES FOR CANADIAN SOCIAL SCIENCE AND  
HUMANITIES JOURNALS**

by

Lindsay Lynch  
Bachelor of Science (Honours), Queen's University, 2003

PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF  
THE REQUIREMENTS FOR THE DEGREE OF  
MASTER OF PUBLISHING

In the  
Faculty of  
Arts & Social Sciences

© Lindsay Lynch 2009  
SIMON FRASER UNIVERSITY  
Spring 2009

All rights reserved. This work may not be  
reproduced in whole or in part, by photocopy  
or other means, without permission of the author.

# Approval

**Name:** Lindsay Lynch  
**Degree:** Master of Publishing  
**Title of Project:** Measuring Online Journal Usage: Issues for Canadian Social Science and Humanities Journals

**Supervisory Committee:**

---

**Dr. Rowland Lorimer**  
Senior Supervisor  
Professor and Director  
Master of Publishing Program

---

**Dr. John Maxwell**  
Supervisor  
Assistant Professor  
Master of Publishing Program

*By written/email consultation from Denmark*

**Dr. Pernille Bjørn**  
Industry Supervisor  
Assistant Professor  
IT University of Copenhagen

**Date Approved:**



SIMON FRASER UNIVERSITY  
LIBRARY

## Declaration of Partial Copyright Licence

The author, whose copyright is declared on the title page of this work, has granted to Simon Fraser University the right to lend this thesis, project or extended essay to users of the Simon Fraser University Library, and to make partial or single copies only for such users or in response to a request from the library of any other university, or other educational institution, on its own behalf or for one of its users.

The author has further granted permission to Simon Fraser University to keep or make a digital copy for use in its circulating collection (currently available to the public at the "Institutional Repository" link of the SFU Library website <[www.lib.sfu.ca](http://www.lib.sfu.ca)> at: <<http://ir.lib.sfu.ca/handle/1892/112>>) and, without changing the content, to translate the thesis/project or extended essays, if technically possible, to any medium or format for the purpose of preservation of the digital work.

The author has further agreed that permission for multiple copying of this work for scholarly purposes may be granted by either the author or the Dean of Graduate Studies.

It is understood that copying or publication of this work for financial gain shall not be allowed without the author's written permission.

Permission for public performance, or limited permission for private scholarly use, of any multimedia materials forming part of this work, may have been granted by the author. This information may be found on the separately catalogued multimedia material and in the signed Partial Copyright Licence.

While licensing SFU to permit the above uses, the author retains copyright in the thesis, project or extended essays, including the right to change the work for subsequent purposes, including editing and publishing the work in whole or in part, and licensing other parties, as the author may desire.

The original Partial Copyright Licence attesting to these terms, and signed by this author, may be found in the original bound copy of this work, retained in the Simon Fraser University Archive.

Simon Fraser University Library  
Burnaby, BC, Canada

## **Abstract**

Features embedded in the technology of online publishing provide an opportunity to redefine readership and measure usage. Such changes are particularly important for Canadian social science and humanities journals that compete for funds from the Social Sciences and Humanities Research Council of Canada (SSHRC)—competing journals are required to demonstrate use of their product. This report explores the potential utility of web analytics for SSHRC’s evaluation of journal usage by unpacking the issues involved in the process of measuring the usage of online journals. In this context, current methods for tracking, collecting, and analysing web usage data are reviewed and the meaning of some common web usage metrics is explored. The findings of this report suggest that implementing standards for collecting, processing, and reporting usage data will support journals in their preparation to apply for funding and will increase the fairness of the competition.

**Keywords:** scholarly journals; electronic publishing; online; web analytics; measure; usage metrics

**Subject Terms:** Scholarly publishing; Scholarly electronic publishing; Scholarly periodicals—Subsidies—Canada; Scholarly electronic publishing—Subsidies—Canada; Electronic journals—Use studies; Electronic publishing—Statistics

## **Dedication**

To my parents, Len and Karen, for your patience and support.

## **Acknowledgements**

I would like to thank Dr. Rowland Lorimer for introducing me to the world of Canadian social science and humanities journals and for providing tremendous guidance and support throughout the course of this work. Thanks to Dr. Pernille Bjørn for helping shape my ideas and to Dr. John Maxwell for questioning them. I am especially grateful to Michael Felczak, Marilyn Bittman, and Anne Carscallen, of the *Canadian Journal of Communication*, for providing me with data and insight about their journal. To the ATIC Lab, including Dr. Guenther Krueger, Dr. Ellen Balka, and especially to my ‘thesis buddy,’ Nina Boulus, for providing the space and camaraderie that facilitated much of this work. Many thanks to Jo-Anne Ray for being omnipotent, and finally, a special thanks to Christopher Savage for steering me back on track whenever I lost my way.

## Table of Contents

<b>Approval</b> .....	<b>ii</b>
<b>Abstract</b> .....	<b>iii</b>
<b>Dedication</b> .....	<b>iv</b>
<b>Acknowledgements</b> .....	<b>v</b>
<b>Table of Contents</b> .....	<b>vi</b>
<b>List of Figures</b> .....	<b>viii</b>
<b>List of Tables</b> .....	<b>ix</b>
<b>Introduction</b> .....	<b>1</b>
<b>Chapter 1: Canadian Scholarly Journals in Context</b> .....	<b>5</b>
Purpose and Value .....	5
The Canadian Market and the Need for Government Support.....	6
Subscriptions as a Traditional Measure of Readership .....	7
<b>Chapter 2: Technology and the Changing Nature of Journals</b> .....	<b>9</b>
The Nature of Online Journals .....	9
Information-Seeking Behaviour .....	10
<b>Chapter 3: Web Usage Data</b> .....	<b>14</b>
The Nature of Web Usage Data .....	15
Collecting Web Usage Data .....	17
Web Server Logfiles .....	17
Page Tagging .....	19
Preparing & Processing Data .....	20
Removing Erroneous Data and Treating Missing Values .....	20
Reconstructing User Activities .....	24
Implications .....	30
<b>Chapter 4: Evaluating Journal Use</b> .....	<b>33</b>
What Can be Measured? .....	34
Hits and Files .....	35
Sessions .....	35
Pages Viewed .....	37
Duration .....	39
Return Visits .....	39
Other Issues .....	42
Journals as a Delivery Vehicle .....	42

Reporting Period.....	42
<b>Conclusions and Recommendations.....</b>	<b>46</b>
Reference List .....	49



## List of Figures

- Figure 1. Total monthly visits to the *Canadian Journal of Communication Online* between January and August 2007, based on 5- and 35-minute Webalizer time-out configuration settings. ....30
- Figure 2. Frequency of total monthly visits and unique visitors to the *Canadian Journal of Communication Online* between January and August 2007.....43

## List of Tables

Table 1. Usage metrics for the <i>Canadian Journal of Communication Online</i> during the month of January 2007 as prepared using Webalizer and phpMyVisites analytics software. ....	31
---	----

## Introduction

This report stems from my involvement with the Canadian social science and humanities (SSH) journal community during and after an internship with Dr. Rowland Lorimer, President of the Canadian Association of Learned Journals (CALJ). In the summer of 2006 we conducted a survey of Canadian SSH journal editors (both editors-in-chief and managing editors, where possible) on their attitudes toward online and open access journal publishing. This study was commissioned by the Social Sciences and Humanities Research Council of Canada (SSHRC), which was, at the time, operating a journal funding support program known as Aid to Research and Transfer Journals (ARTJ). Our findings were published in a report, entitled *Augmenting Print: Planning for Online Journal Publishing by Social Sciences and Humanities Journals in Canada*, which contained ten recommendations to SSHRC for planning future changes to the ARTJ program (see Lorimer, Lynch, & Provençal, 2006).

SSHRC's involvement as a major stakeholder in the success of this journal publishing community suggested the need for a more detailed examination of its policies and their impact on journals. SSHRC funding has come to be critical to journals eligible for support. Without financial assistance from SSHRC, many valuable journals would not be financially viable.

After the release of *Augmenting Print*, SSHRC initiated a pilot program for funding online open access journals, Aid to Open Access Journals (AOAJ)—the first instance of funding for journals operating outside of a traditional print-based subscription

model. In 2008 SSHRC launched its most recent iteration of aid for scholarly journals, the aptly titled Aid to Scholarly Journals (ASJ). This new program merges ARTJ and AOAJ to form a single competition that is open to eligible journals, “regardless of business model or distribution format” (SSHRC, 2008). These new online journal initiatives present particular challenges for the measurement of readership, which is, in effect, an indicator of demand and, arguably, the contribution such journals make to the community. Prior to the emergence of ASJ, SSHRC (SSHRC, 2007b: n.p.) considered subscriptions “to be an accurate proxy for regular readers.” For purposes of grant eligibility, it set a minimum expectation of 200 subscribers. Since this proxy cannot be applied to open access journals, because by definition they do not have subscribers, the AOAJ program then asked for a demonstration of “at least 250 regular readers” (SSHRC, 2007a: n.p.). ASJ is now requiring that journals “have a minimum of 250 regular readers, as demonstrated through a detailed web-usage report or other verifiable documentation, such as a list of subscribers” (SSHRC, 2008, n.p.).

This attempt to create a parallel metric among print, online, and open access journals in a world where a rich set of measures of access behaviour can be collected and reported led to the question this report seeks to address: What are the issues inherent to measuring online journal usage? This question is both timely and relevant because SSHRC is in a transition period with respect to its journal funding policies, just as journals are in transition to making the best use of print as well as the online environment. The intent of this report is to provide insight for stakeholders that will foster evidence-based decision-making regarding an issue that will have a significant impact on the scholarly journal publishing community in Canada. This report does not

attempt to solve the problem of what to measure, but rather discusses the sources of uncertainty and subsequent complexities involved in the process of tracking, collecting, and analysing web usage data. Although these activities are facilitated by widely available software, the implications of using the resulting metrics to compare and rank journals are highly complex.

One other issue not addressed in this report is the potential utility of Google PageRank. PageRank employs an algorithm that determines the popularity of a web site based on the number and type of other web sites that link to it. This approach to value measurement is similar to the approach used by Thomson Reuters ISI to calculate journal impact factors. Such approaches to evaluating the utility or value of an information source go beyond the notion of readership, and as such are beyond the purview of this report.

Chapter 1, *The Value of Scholarly Journals*, lays down the foundation for the empirical work by identifying the purpose and value of scholarly journals. Background information on the market for scholarly journals in Canada helps contextualize the decision by the federal government to provide these journals with various forms of support. Readership is introduced as a traditional measure of demand for scholarly journals. Chapter 2, *Technology and the Changing Nature of Journals*, identifies the essential difference between print and online journals and examines how this difference has changed the nature of journal use, with a particular focus on information needs of researchers. These changing habits warrant a new approach to measuring readership. Chapter 3, *Web Usage Data*, introduces the technical aspects of collecting and measuring web-based journal usage, the process of web analytics. As well, the nature of web usage

data is discussed as it has implications for how usage data are interpreted. Chapter 4, Evaluating Journal Use, explores some of the metrics that can be generated from the data collection methods discussed in Chapter 3, including what each metric really says about journal use. The report concludes with some final thoughts and recommendations for SSHRC.

# **Chapter 1: Canadian Scholarly Journals in Context**

## **Purpose and Value**

Scholarly journals act as the official record for original research that is intended to be accessible to all other researchers to serve as a knowledge foundation on which further research is based. This is, however, not their sole purpose. Such journals also function as a means of connecting people (researchers, students, practitioners, policy makers, etc.), the importance of which should not be underestimated. According to Lorimer & Maxwell (2007: 176)—who were writing specifically about Canadian scholarly journals—it is important, especially for new scholars, to know “what research is going on in Canada, who the participants are in the Canadian research community, what questions they ask, where they work, and so on.” Finally, as an artefact of the tendency for journals to emerge around disciplines, scholarly communities, geographical regions, and scientific approaches, they end up structuring knowledge so that it can be more easily retrieved.

By fulfilling these obligations, scholarly journals “play an essential role” in scientific communication, both by participating in “the development and distribution of knowledge in their respective societies” and by acting as “instruments of recognition, legitimation, acknowledgment, dissemination, and enhancement of scientific heritage” (Boismenu & Beaudry, 2004: 346).

## **The Canadian Market and the Need for Government Support**

Canadian SSH journals participate in the value system described by Boismenu and Beaudry by promoting the visibility of Canadian research and Canadian researchers (Lorimer & Maxwell, 2007). However, journals that publish such work operate in a particularly challenging publishing environment. Research that is specific to the Canadian context has increased value for Canadian audiences, but this specificity may reduce the demand from international audiences for such work. The common markets for journals in general and in rank order are institutional libraries, individual scholars, students, sometimes professionals, and society members (Lorimer & Lindsay, 2004). The common markets for Canadian SSH journals are a subsection of these general markets. Since the domestic market is small, and the particular sub-disciplines even smaller, it is challenging to operate a journal that may be in high demand by its audience, but whose audience is too small to financially sustain the business of publishing.

A typical Canadian SSH journal tends to have around 400 subscribers (Lorimer & Maxwell, 2007). Although this number may sound low, it may actually represent a large portion of the market share of a journal's target audience, especially considering that institutional subscriptions make up a large portion of subscribers. Thus, in order to recoup the costs associated with producing a journal for such a small audience, subscription rates would have to be much higher than audiences are able or willing to pay (Lorimer & Maxwell, 2007). These circumstances are bleak enough that, as Lorimer and Maxwell (2007: 176) point out, without subsidies, "few Canadian social science and humanities (SSH) journals would exist".



The precariousness of Canadian publishing in general and Canadian scholarly journal publishing in particular is not a new phenomenon, nor is it likely to change in the short term. Hence, the federal government supports SSH journals through SSHRC's journal funding programs mentioned in the Introduction. SSHRC's commitment to supporting SSH journals reflects the Council's belief that high quality "scholarly journals are a primary tool for fostering intellectual debate and inquiry" (SSHRC, 2008). SSHRC justifies funding these journals by acknowledging that work published in them, work built upon their contents, and the communities of research that grow out of them return the investment as public goods.

In providing financial aid, SSHRC is faced with the difficult task of doling out limited funds to a seemingly unlimited group of needy journals. In order to avoid the problem of funding journals that no one wants to read, SSHRC requires journals to prove that there is reasonable demand for their product. In terms of serial publications such as newspapers, magazines, and journals, readership is the term that is used to indicate demand.

### **Subscriptions as a Traditional Measure of Readership**

The readership of print-based journals was traditionally measured in terms of subscriptions: the number of subscribers was assumed to be an indicator of how many people used the journal. The validity of this assumption was related to the notion that a subscriber's willingness to pay money for the journal meant that the journal's content was somehow of value to the subscriber. However, many journals come bundled with membership in a professional association. In such cases, the number of subscriptions may not be synonymous with how well the publication meets its users' needs.

Subscriptions of individual scholars are best viewed as a means of ensuring that they have ready access to the contents of a journal, should they have a need for information contained within it and a knowledge of where it can be found. Institutional subscriptions can be viewed as a means of ensuring that the scholarly community has somewhat ready access to the contents of a journal should it need it. (Libraries provide the added value of finder's aids in locating relevant documents.) This "just-in-case" interpretation of subscription volume implies that subscriptions are an inflated indicator of the potential of a journal to meet information needs of its audience. This subscription/use pattern breaks down when journals are unbundled from professional membership (surely you get a truer reading of who is actually using the journal). It also breaks down as subscription prices rise. And as we will see, subscriptions as an indicator of readership breaks down in an online environment and is completely non-existent in an open access environment.

Fortunately, features embedded in the technology of online publishing provide an opportunity to redefine readership and measure usage. The rest of this report explores this possibility by unpacking the issues and complexities involved in the process of measuring online journal usage, beginning with an exploration of how the technology of online publishing has affected the nature of journal use.

## Chapter 2: Technology and the Changing Nature of Journals

### The Nature of Online Journals

Accompanying the transition to an online format is a shift in the journal's basic nature. While the essence of the journal is still the same—it remains an entity that has the goal of collecting and vetting scholarly content for the purpose of contributing to knowledge—the journal's identity is no longer represented by a collection of printed articles bound by a cover. Instead, it is represented by a web site, with various links, features, services, and collections of information. This chapter articulates what these changes are and how they impact journal use, because understanding how use has changed can inform decisions about how to evaluate it.

Mackenzie Owen conceptualizes the electronic<sup>1</sup> journal as “a construction that results from the way a social entity (researchers) utilize technical possibilities (the outcome of the process of digitization) when they disseminate their knowledge in the form of recorded information” (2007: 14). The essential difference between print and online journals is that they are constructions based on two distinct technical possibilities—print media and digital media. Regardless of format, the interaction between the user and the journal's content is mediated by some form of technology—either a book (for print journals) or a computer and the associated hardware and software required to render a web page (for online journals) that can, if the users wishes, be

---

<sup>1</sup> The term *electronic journal* may refer to any digital journal, not only those published online (e.g., CD ROM)

printed. Of course, there is a much wider set of differences that derive from this initial distinction.

In discussing the technical possibilities presented by digitization to scholarly journals, Mackenzie Owen (2007: 10) suggests that, at present, the electronic journal article “as a communicative form for reporting on research and for disseminating scientific knowledge...remains a digital copy of the printed form.” For the most part, this is true, with notable exceptions in which sound and moving images are deployed in the service of portraying reality. For Mackenzie Owen, the differences between print and electronic journals are “at the level of the infrastructures developed by publishers” (2007: back cover), resulting in “the creation of new and highly sophisticated mechanisms for retrieval, linking, access control, delivery and licensing” (2007: 10).<sup>2</sup> Since the nature of the differences between print and online journals with respect to usage relates to access and retrieval, we might expect that these changes will have an impact on the way users access and retrieve information produced by journals.

## **Information-Seeking Behaviour**

As journals are an information product, they may serve the information needs of their audiences in a number of ways. To get a better picture of what those needs are, it is helpful to look at what information-seeking behaviours researchers exhibit during the research cycle. Although information seeking is a stochastic and highly individual process (Case, 2002), the practice of information seeking can be broken down into a series of activities which were initially characterized by Ellis in 1989 as starting,

---

<sup>2</sup> Interestingly, this leaves out submission management and production, which is now *de rigeur* in online journals.

chaining, browsing, differentiating, monitoring, and extracting. Since Ellis' model was developed before the emergence of the World Wide Web, more recent work has been done to account for this important development, thus expanding Ellis' model to include four additional features—accessing, verifying, networking, and information managing (Meho & Tibbo, 2003). These activities make up three general stages of information-seeking known as searching, accessing, and processing (Meho & Tibbo, 2003).

According to Meho and Tibbo (2003: 571), Ellis' categorization “suggested that information retrieval systems could increase their usefulness by including features that directly support these activities.” All journals can function as information retrieval systems, but web-based journals have the potential to support these information-seeking activities differently than paper-based journals. Specifically, the combined potential of digital media and the network capabilities of the Internet have resulted in a number of value-added features that support the information-seeking activities of a journal's users in ways that were not possible with the paper-based system.

To begin with, there is a greater emphasis on indexing and searchability in a networked environment, so the ability to find and retrieve information is strengthened both within the journal and in the universe of all journals. Many journal web sites have a built-in search functionality that facilitates directed searching. Users can search a journal's archive by author, title, publication year, key words, etc. Users can also browse the table of contents of current and back issues as a form of semi-directed searching. The act of chaining—following a chain of citations from one work to another—is facilitated by journals that provide a linking service from the works cited in an article to the

electronic version of those works hosted by other vendors (e.g., through CrossRef).

Tenopir et al. (2003) have reported an increase in this type of behaviour.

Online journals facilitate the ability of users to monitor of the emergence of new content by providing services such as email updates and RSS feeds (e.g., table of contents of each issue; notifications when new articles are published that meet pre-specified search criteria). In accessing a full text article, users may have access to a “reading tool,” which allows them to view details about the author(s) and metadata about the article, allowing them to differentiate—to use “differences between sources as filters on the nature and quality of the material examined” (Ellis, 1989: 178)—between sources (see Open Journal Systems software <pkp.sfu.ca>). Processing the relevance of an information source—assessing the general relevance followed by close analysis of specifically applicable content of a particular document—is facilitated by the digital nature of documents on the Web, which are fully searchable. Thus, users can scan documents in order to extract key material very quickly (by searching the full text of the article for key words or phrases). Verification of study results is now being facilitated by journals that provide access to datasets. This practice is more common for natural sciences and medicine than for social sciences, where the nature of the data is such that it may not be ethical to make it public.

Communication between scholars for the purposes of networking and providing/receiving feedback may be supported by the provision for review and commentary online, although these activities are restricted by the degree to which the content serves as a record rather than being engaged with on reading as in a blog.

Decisions about where/how to access sought-after material are supported not only by the vast resources of Google and other search engines including Google Scholar, but

also by the collaboration of libraries and vendors who provide aggregated access to a large number of journals that libraries could not otherwise afford or physically store in print. The ‘Where can I get this’ feature that many libraries have embedded in their catalogues allows users to see their options for accessing a particular article and to choose the one that is most convenient.

Information managing is also facilitated by journals that provide a citation for the article being viewed, and many journals allow users to export the citation information to reference management software (e.g., Reference Manager, EndNote). The portability and storability of electronic documents is a feature that can be leveraged by researchers to manage their personal collections.

With all of these new features and ways of using the journal, it makes sense to re-evaluate the approach to measuring readership. The next chapter deals with the question of how the features embedded in the technology of online publishing can be leveraged to obtain data about online journal usage.

## Chapter 3: Web Usage Data

Delving into usage beyond the empirical elements outlined in Chapter 2 demands a consideration of the technical aspects of web sites, web usage, and how web usage data are collected. The practice of collecting, measuring, and analysing web usage data is called *web analytics*. This chapter explains how web analytics techniques work, thereby providing a context for thinking about what can be measured and about the meaning behind the measures. Expressed in different terms, this chapter unpacks the notion of web analytics and begins to explore the implications of drawing conclusions about behaviour based on electronic usage data.

There are two main literatures that inform this report. Academic literature, particularly from the field of library and information science, and business literature, often in the form of white papers and web sites prepared by consulting firms. The former literature deals with questions about information-seeking behaviour and how to provide and evaluate information systems that support that behaviour. The intention of the latter, it appears, is to help individuals evaluate and improve their web sites with the aim of increasing the success of their business. The underlying assumption that is common among the business literature is that a well-made web site will increase the chances of business success<sup>3</sup>. The way to improve a web site is to track how customers use it and to extract from the usage patterns clues about what might be preventing customers from

---

<sup>3</sup> Business success is generally measured in terms of customer conversion (Spiliopoulou & Pohle, 2001). Customer conversion refers to the desired trajectory of a site visitor from the entry page to through to the point of purchase. A converted customer is one who follows through to purchase a product the site is selling.



achieving the site's goal (i.e., making a purchase). In that sense, web analytics is a process of continuous testing and reflecting about what drives customers. Site operators can then make informed decisions about how best to modify the site to match customer needs.

The purpose of obtaining usage metrics in the context of the SSHRC journals competition, however, is not to improve the business success of journal web sites; it is, rather, to provide a basket of measures that fairly reflects the functioning of all applying journals. This goal can be broken down into two problems: 1) establishing basic eligibility and 2) assessing the quality of the performance of those journals that are deemed eligible. Thus, interpretation of usage metrics in this context will have a different meaning than when site operators interpret usage metrics for the purpose of tracking the improvement of individual web sites.

## **The Nature of Web Usage Data**

The interaction between the journal user and the contents of a journal's web site is mediated by a computer network model known as client-server software architecture. This model "describes the relationship between two computer programs in which one program, the client, makes a service request from another program, the server, which fulfills the request" (Client-server, n.d.). In the case of online journals, the client is the web browser (e.g., Mozilla Firefox) that renders the web page (e.g., a full-text article) that has been returned by the server that hosts the web site.

The practice of web analytics does not involve the direct observation of human behaviour. Instead, it is the practice of observing a record of electronic transactions

between the client program (web browser) and the web server(s) that electronically administer(s) a web site. The client may represent an individual or groups of individuals (Nicholas et al., 2006), and there may be many clients operating in a single terminal by a single user (e.g., if an individual has several browsers running simultaneously). And since it is impossible to know, with any degree of certainty, exactly who is operating the computer that invokes the browser, the concept of user in this context is ambiguous. To illustrate this point, consider a computer terminal in a library. Over the course of a day, many different individuals may access the terminal, conducting searches and downloading articles. But there is no mechanism to distinguish between these individuals (unless they are forced to log into the system with a user ID and password—but this is not often the case), so the record of interactions between that terminal and the web sites it has accessed will appear as a single user. And inasmuch as the user is ambiguous, so too are the users' intentions. For example, an individual may click accidentally on a particular link, may click more than once on the same link, the computer may freeze, causing the user to quit a session by closing the browser and rebooting the computer, etc. Such ambiguities are inherent in web analytics data. If acknowledged as a construct with inherent ambiguities, web analytics is a useful tool, but the data must be interpreted with care and an awareness of these limitations.

In addition to the ambiguous nature of web analytics data, capturing human behaviour by means of such a proxy involves multiple translations of information. That is, behaviour is captured as electronic transactions between browsers and servers, and these transactions are recorded as lines of text. These lines of text are then processed and analysed to remove instances of unintentional use (see below for explanation). Then

information on patterns is extracted and interpreted as human behaviour. Each translation is prone to errors and uncertainties that reduce the extent to which the data reflect the true behaviour of the user community (Spiliopoulou, Mobasher, Berendt, & Nakagawa, 2003). If decisions are going to be based on these data, then decision makers are compelled to understand how the data are collected and constructed.

In order to illustrate the points at which uncertainties are introduced, it is helpful to think of web analytics as a process that involves three phases: 1) data collection, 2) data preparation and processing, and 3) data analysis and interpretation. This chapter covers the first two phases; the third will be addressed in Chapter 4.

## **Collecting Web Usage Data**

Uncertainties are introduced at the data collection level because of the ambiguous user problem discussed above. Instead of directly observing human behaviour, the observable phenomena are interactions between client software and web servers. The way these interactions are captured and recorded has an influence on data quality. Logfile analysis and page tagging are the two most common methods for capturing and recording information about these interactions (Web analytics, n.d.). Both methods involve similar techniques for storing information about the electronic transactions between client and server. The main difference is the mechanism by which they are collected.

### **Web Server Logfiles**

Client-server transactions are automatically and faithfully recorded as an embedded part of the web server system. Each transaction is recorded as a single line of text, and a collection of these lines is called a logfile. The individual lines of the logfile

are composed of several pre-defined metadata fields (“tokens”). Conceptually logfiles are equivalent to a guest book, only instead of visitors’ names, other pieces of information about that interaction are recorded. The standard format for logfiles is called Common Log Format (CLF) (W3C, 1995). The box below contains a sample entry from a logfile in this format. Each log entry in this format contains the following information:

192.168.45.13 - - [24/May/2005:11:20:39 -0400] "GET /mypage.html HTTP/1.1" 200 117						
a	b	c	d	e	f	g

- a) Remote hostname (domain name or IP number if the name is not available)
- b) Identity information reported by the client (if the client has enabled this feature)
- c) User ID (if the request was for password protected document)
- d) Date and time of the request
- e) Exact URL of the page requested by the client
- f) Status of the request (W3C return codes)
- g) Size (in bytes) of the object returned to the client (W3C, 1995).

Although it is beneficial that the server system automatically records this information, which means no additional effort is required to collect it, it should be noted that this method was originally designed for the purpose of tracking server system performance. For instance, the logfiles contain information on failed requests that may be caused by server inadequacies. Kaushik articulates the inherent deficiency of using web server logs for anything other than tracking system performance by noting that “web logs were built for and exist to collect server activity, not business data” (2008). The system was designed to tell web site administrators when the server had problems fulfilling requests, not to tell them who was requesting what and why.

An additional limitation to this method of data collection is that browsers have built-in mechanisms that are meant to reduce the burden on servers (i.e., minimize the number of requests being sent to the server). This mechanism is known as a *cache*. A browser's cache stores information from web pages that have been delivered by the server so that future requests for those pages are fulfilled by the cache. This mechanism prevents a request from being sent to the server, which means those repeated requests are not recorded in the logfile. As a result, there are limitations about what conclusions can be drawn from such data.

### **Page Tagging**

Page tagging involves the use of JavaScript to help track which web pages (pages that make up a web site, not pages from individual articles) are being accessed and used. The facilitating script is embedded on each page of a web site, and every time a web page is loaded, the script is run. The script sends the same kind of information that a web server logfile collects to a remote (third-party) server, where the transactions are recorded in a logfile or database. Since the script, which is the signal indicating use of a web page, is attached to the page rather than the server, it is run every time a page is loaded, regardless of where it came from, so caching is not an issue for the data set. The script collects essentially the same information about the client as the web server logfile, with the potential to gather additional information about the client browser, operating system, screen resolution, etc. Such additional pieces of information are beneficial for analysts interested in optimizing the web page design and functionality for its users, but have little to do with gauging web site activity.

These are two common methods for collecting web usage data, and are by no means the only methods that exist. For the purposes of this report, however, they represent sufficiently the complexities and issues that are inherent to measuring web usage in general and usage of online journals in particular. Ultimately the data captured by these methods represent nothing more than the requests for particular web pages. An explanation of by whom and for what reason these requests were made is not addressed by either methodology. Answers to these questions must be inferred by analysts based on assumptions about how web users behave.

## **Preparing & Processing Data**

Uncertainties in the data are uncovered at the data preparation and processing level where decisions have to be made about the quality of the data and how to address the shortcomings. Failure to adequately prepare the data renders them useless (Spiliopoulou et al., 2003). As such, in the context of a funding competition, it is imperative that applicant journals submit metrics derived from adequately prepared data.

The data preparation process involves repairing erroneous data, treating missing values, removing “requests that do not reflect human navigation behaviour, i.e., requests from robots and other software agents,” and reconstructing the “activities of each individual user during each visit” to the site (Spiliopoulou et al., 2003: 173). Each of these tasks is discussed in greater detail below.

## **Removing Erroneous Data and Treating Missing Values**

Project COUNTER is an industry-driven initiative with the objective of providing standardized usage data to libraries to help them inform purchasing decisions (see

<http://www.projectcounter.org/>). This organization provides instructions for journal vendors that wish to be COUNTER-compliant on how to process their web usage data and what key metrics to present to libraries and other interested stakeholders. The issues that COUNTER raises provide a useful framework for explaining the complexities of processing raw usage data.

### **Valid Requests**

COUNTER distinguishes between intended and unintended usage, the former being of sole interest to the project. Thus “all requests that are not intended by the user” must be removed from the dataset (COUNTER, 2008b: 35). Robots, spiders, and other software agents crawl the web for indexing purposes. This activity invokes the same server request mechanism as the browser request triggered by human users. Since these requests do not reflect the activities of individuals looking at a journal, they need to be removed from the dataset before the analysis can occur. In addition, protocols that are meant to increase the efficiency of the Web generate requests to the server that do not reflect intended usage. For instance, some search engines perform a “prefetching” activity when they return the list of search results (COUNTER, 2008b: 36). The first few results in the list are assigned a <link> tag, which triggers the browser to “fetch the pages behind the results and load them into the browser cache” (COUNTER, 2008b: 36). If such links are to full text articles, then this will show up as a successful request for a full text article. If the individual operating the browser then intentionally requests the full text article, it will appear as a second request in the log. (See COUNTER, 2008b: 36-37, for further explanation.)

COUNTER code of practice also requires that “records generated by the server together with the requested page (e.g. images, gif’s, style sheets (.css)) should be ignored” (COUNTER, 2008b: 35). Since web pages may be composed of individual elements, such as images, videos, etc., requests for these individual components are recorded in the logfiles. However, in the analysis we are not interested in this fine level of detail, but in the request for a page as a whole, which means these items need to be filtered from the dataset.

Another source of invalid requests is the problem of double-clicks. As previously mentioned, all transactions between the client browser and the web server are recorded. Requests are generally initiated by the user who clicks on a hyperlink using the mouse, and for a variety of reasons that most readers are familiar with, it is not uncommon for Internet users to click more than once on a link when trying to load a page. The server does not discriminate between clicks that are intended to retrieve a page, accidental clicks, and impatient clicks; thus all such clicks are recorded as requests for that particular page. Without observing the actual human behaviour, it is not possible to know with 100 percent certainty which clicks fall into these categories. Nonetheless, it is important to address this phenomenon when analysing web usage data. Otherwise, it is possible to end up with an inflated picture of reality.

COUNTER’s suggested solution to this problem is a heuristic<sup>4</sup> method whereby a pre-determined window of time between requests for the same http-link is used to determine whether or not a double-click has occurred. They have designated a time of 10 seconds as the window between the first and second click—any requests for the same

---

<sup>4</sup> A heuristic is an approach to solving a problem for which there is no definitive algorithm, usually based on trial and error.



page that are spaced more than 10 seconds apart are considered two separate, valid requests (COUNTER, 2008b: 35).

It is well known that Internet users will often click refresh or back in the midst of waiting for a web page to render, or immediately after it has rendered. This type of behaviour introduces an additional level of uncertainty into web usage data, because, as with the example of double-clicks, assumptions must be made about the user's actions and intentions. The issue is particularly pertinent in the case of online journals, where many publish both HTML and PDF versions of the same article. Users may be inclined to change their minds as to which version they want to view (especially since PDFs take longer to download than HTML files). COUNTER addresses this issue by requiring that "requests by one and the same IP/username/session- or user cookie for one and the same PDF should be counted as a single request if these multiple requests occur within a 30 seconds time window" (COUNTER, 2008b: 35-6). In such cases, vendors are advised always to "remove the first and retain the second" request (COUNTER, 2008b: 36). The same strategy is suggested when one and the same HTML article is requested, but within a shorter (10-second) time window and when a request for an HTML article is followed by a request for the same article in PDF form (COUNTER, 2008b: 36).

### **Successful Requests**

Unsuccessful requests to the server constitute another source of erroneous data. COUNTER recommends that they be removed from the dataset, and that only successful requests be counted (COUNTER, 2008b: 35). The status of a request is identified by the

NCSA return codes<sup>5</sup> that appear in each line of the logfile. COUNTER defines successful requests as those with return codes 200 and 304. The 200 code signifies that the request was successfully received. The 304 signifies that the item requested can be retrieved from the browser's cache because it has not been modified since the last time it was requested (COUNTER, 2008a). COUNTER guidelines recommend that all other return codes be discarded from the dataset.

Because the objectives of SSHRC's evaluation may differ from those of COUNTER's, there may be some merit in examining the code set to determine if other results are of interest. For instance, unsuccessful requests that are the result of a server error may be worth reporting since these requests reflect interest in the journal's content that otherwise would have been fulfilled. Of course, any such decisions will require thoughtful examination of the implications and how they align with the objectives of the competition.

### **Reconstructing User Activities**

Once the dataset has been cleaned up, the next task is to reconstruct the user activities so that meaning can be extracted. This process is based largely on assumptions about user behaviour. The first step in reconstructing user activities is to assign the series of electronic requests to individual users, and there are few ways this can be done. The multitude of individual requests must then be separated into visits—a process known as *sessionization*.

---

<sup>5</sup> See <[http://www.projectcounter.org/cop\\_books\\_appendix\\_d.html](http://www.projectcounter.org/cop_books_appendix_d.html)> for further explanation.

## **Identifying Unique Visitors**

Web server logs are “the primary source of data in which the activities of Web users are captured” (Spiliopoulou et al., 2003: 172). However, the information contained in the logs is incomplete, since “the requests they contain cannot be uniquely assigned to the individual that has performed them” (172). Because there are several ways that a user can enter a web site, identifying unique visitors is a significant challenge. For instance, Nicholas et al. (2006: 1350) identify “via trusted proxy server, as a society member, location such as a university, IP address, username and password” as a few of the ways that users will enter a journal’s web site.

Of course, the easiest way to keep track of individuals is by user authentication. This technique requires users to be registered with the site and to have a user name and password that will allow them access to the content. This is often the course of action taken by subscription-based online journals as a means of restricting access to subscribers. However, it is worth noting Nicholas et al.’s (2006) findings that only 10 percent of subscribers enter web sites via their user name and password. The behaviour reported by Nicholas et al. (2006) may be explained by the fact that a large proportion of individuals with personal subscriptions may also have access to the journal via an institutional subscription, which, if working from the campus, will likely have automatic access through the institutional network. Moreover, it is not considered good practice to restrict content in this manner, as such access barriers are known to reduce use overall, which contradicts the idea of maximizing dissemination of knowledge (Bittman, Lynch, & Pauls, 2007: 75).

Cookies are another method that can be used in both logfile analysis and page tagging as a means of identifying unique visitors. A cookie is a small bit of text sent from a web server to the requesting client browser (HTTP cookie, n.d.). For each subsequent request from the client browser, the bit of text is sent back to the server, allowing the server to recognize the client as having made previous requests. This is the technology that allows consumer web sites to provide virtual shopping carts to users so that they can store items for later purchase while they browse the site.

Cookies can be assigned by the server of the web site of interest or they can be assigned by a third party. The former type are known as *session cookies* because they only track the activities of a client browser for the duration of that session on that particular web site. The latter type, referred to as *third-party* or *user cookies*, stay with the client browser as it moves from site to site. This type of cookie is usually assigned by advertisers, whose intent is to track users across multiple web sites. Session cookies are often used for identifying unique visitors in conjunction with logfile analysis (if the server is configured to assign cookies), whereas page tagging often relies on third-party cookies because the service can be outsourced to web analytics firms.

Cookies are used because IP addresses are not always unique to users and may be shared by large group proxies. It should be noted, however, that many Internet users are in the habit of deleting third-party cookies on a regular basis—it is as simple as changing a browser setting. According to Dainow (2005: n.p.), “58 percent of users delete their cookies regularly, with 40 percent deleting them every month.... However, only 1 percent delete cookies set by the site itself -- it is third-party cookies that people are

deleting.” This tendency means that estimates of repeat use could be vastly underreported by journals tracking users with third-party cookies.

The least reliable strategy for identifying unique visitors is by IP address. This method involves using heuristics to identify users after they have visited the site (in contrast to authentication and cookies, where identification occurs before or during the visit to the site). Since the process of associating requests with individuals occurs after the interaction, this method relies on assumptions about user behaviour, and is therefore less certain than the previous two methods.

The uncertainty involved in the process of identifying unique visitors is reflected in the discourse of the web analytics community. The documentation for Webalizer, a popular, free web analytics program uses particular language to refer to the metric that can be obtained from adding up “the number of unique IP addresses/hostnames that made requests to the server” (Barrett, 2008: n.p.). The term used by Barrett (2008) is *sites* (as in the various client sites that have made requests to a server. Barrett (2008: n.p.) cautions that “care should be taken when using this metric for anything other than that.” However, in general parlance this metric is known as *unique visitors*. The reason for cautioning the use of this terminology is that “many users can appear to come from a single site, and they can also appear to come from many IP addresses so it should be used simply as a rough gauge as to the number of visitors to your server” (Barrett, 2008: n.p.).

## Sessionization

The process of identifying individual requests from the same client during the same visit to a web site and grouping them together to reconstruct the user's activities is known as *sessionization*. Grouping the data into visits allows analysts to infer the navigation behaviour of the individual users, which is the point where meaning can be extracted about how the site is being used. Spiliopoulou et al. (2003: 173) note that sessionization "is no trivial task, because the data recorded by a Web server are not sufficient for distinguishing among multiple visits of the same person." The methods used to perform this task can have a significant effect on the metrics derived from the dataset (see Chapter 4).

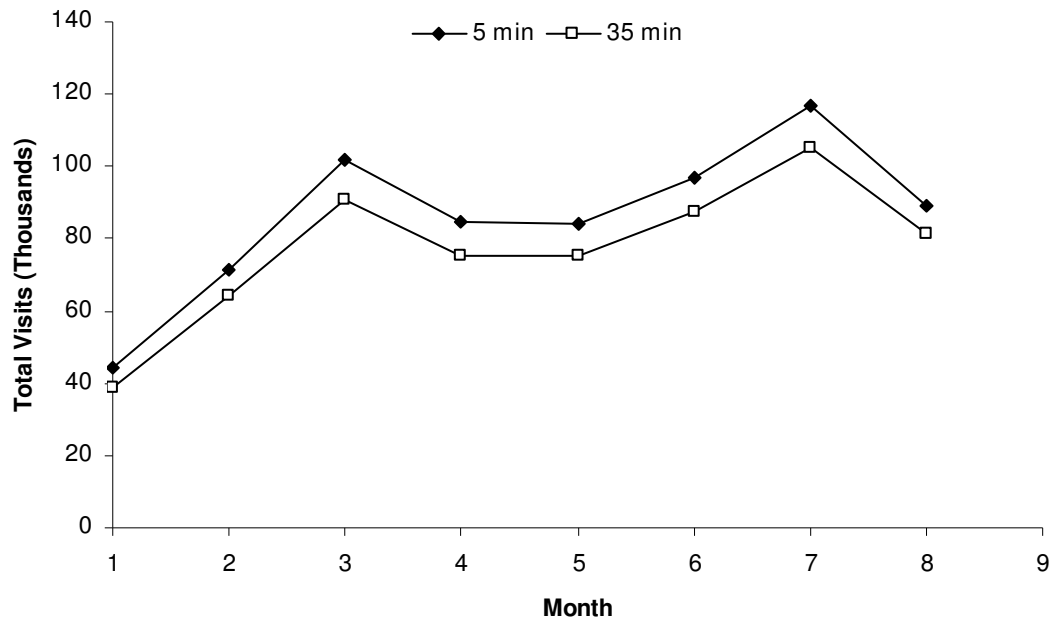
A logfile contains a list of requests to the server in chronological order. This means that requests from one and the same user do not necessarily appear as consecutive lines in the log (since multiple individuals may be using the site simultaneously). So these requests need to be grouped together to reconstruct the activities for that individual.

A session is by definition a time-bound metric. That is, all the activities performed by an individual during a particular visit to a web site from the time they enter the site to the time they leave it are described by the term *session*. Because of the nature of the client-server relationship (i.e., that communication between the two is based on requests), there is no mechanism in place to definitively indicate when a particular session finishes, i.e., when the user is done using the site. "Hardly anyone logs off from a site, they simply leave the site and this is conducted quite anonymously as far as the logs are concerned" (Huntington et al., 2008: 359). As such, guidelines for performing

sessionization are based on assumptions about how much time users generally spend using a web site.

Most commonly, a predetermined interval of time, referred to as the time-out setting (let's call it  $t$ , usually measured in minutes) is used as a cut off when grouping activities into sessions such that any requests that are made more than  $t$  minutes apart are considered to be from separate visits. Huntington et al. (2008) provide a useful review of the literature on the subject of sessionization in which they describe a number of studies where time-out settings are established based on user experiments. Such intervals range from 8.6 to 30 minutes. The default time-out setting for most web analytics software is 30 minutes, although it can be configured to whatever time-out setting is desired.

It is not difficult to see how metrics based on data generated in this manner may be inaccurate and/or easily manipulated. Figure 1 illustrates the impact of using different time-out settings on the total visits metric for the *Canadian Journal of Communication* (CJC). In this example, the time-out settings used were 5 minutes and 35 minutes. There is a difference of more than 11,000 visits for the month of July when the total visits metric is calculated using these two settings.



**Figure 1. Total monthly visits<sup>6</sup> to the *Canadian Journal of Communication Online* between January (month 1) and August 2007, based on 5- and 35-minute Webalizer time-out configuration settings.**

## Implications

Table 1 contains a list of usage metrics for the *Canadian Journal of Communication* that were obtained using two different web analytics approaches. Webalizer (version 2.01) was used to process data from the journal's web server logfiles. This program does not automatically filter for spiders and robots, so presents an inflated picture of usage. The second program, phpMyvisites, is an open source web analytics program that uses page tagging and session cookies to collect usage data. Both programs were configured with 30-minute time-out settings. This table illustrates the potential for

<sup>6</sup> It should be noted that these data have not been filtered for spiders and robots, and so represent an inflated picture of journal use. However, absolute values are not necessary to illustrate the difference between visits calculated using the two time-out settings.



variation between metrics from the same journal when data collection and processing are performed using different web analytics techniques and configuration settings. The meaning of the metrics themselves will be discussed in greater detail in Chapter 4.

**Table 1. Usage metrics for the *Canadian Journal of Communication Online* during the month of January 2007 as prepared using Webalizer and phpMyVisites analytics software.**

Metric	Description	Webalizer	phpMyVisites
Hits	Total number of requests made to the server.	199,504	–
Files	The total number of requests that were successfully fulfilled by the server.	162,017	–
Pages viewed	Total URLs requested that represent a web page as a whole, not its individual components.	100,596	32,816
Sessions (Visits)	Total number of sessions conducted.	39,236	11,939
Sites (Unique visitors)	Total number of unique clients making requests to the server.	13,454	10,152
1-page visit rate	The percentage of clients that requested only one page.	–	64%
Returning visitors	Number of clients that conducted more than one session.	–	1177
Returning Rate	The percentage of unique clients that returned to the web site.	–	12%
Number of visits per visitor	The average number of sessions conducted per client.	–	1.2

This chapter has demonstrated the great variability of web usage tracking and processing methods that are available for use by scholarly journals. The datasets obtained using these methods are analyzed to generate metrics that tell us something about how the journal is being used. These resulting metrics, which will be discussed in the next

chapter, are affected by the choices made regarding the collection and processing methods presented here. For instance, the way the data are filtered and processed will have implications for metrics like number of page views and number of downloads. The way that unique visitors are identified will have implications for metrics like number of unique visitors and the number of repeat visits. Thus, it is important for SSHRC to consider implementing standards for data collection and processing, perhaps by enforcing the use of standard software with specific configuration settings for all journals applying for funding.

## **Chapter 4: Evaluating Journal Use**

There is a sense in which journal reading was completely unexamined in the print world. Familiarity with the research literature relevant to one's topic was expected to be complete, at least until the late 1960s, and that requirement remains, albeit more narrowly defined. Researchers demonstrated familiarity in both the conception of a research question and in citations. But how researchers queried the literature was unknown and undiscussed. With usage able to be traced, we now can attempt to describe how people access journals, to identify the search terms used, to calculate the average time they spend looking, and to identify which and how many articles they tend to access. These attributes divulge what we mean by reading. They also open up questions about how use should be defined and what measures should be collected to support new notions of readership.

What constitutes use of an online journal is a question that needs to be part of a larger discussion than is addressed in this report. This chapter makes a foray into such a discussion, underlining the notion that no single metric is adequate enough to present a reasonable picture of journal usage. It might be important, for instance, to report separately the behaviour of single-visit users from those that visit multiple times, viewing both as important descriptors of journal usage. What we consider 'legitimate' use might be purely instrumental, involving only those usage events that contribute to the practice of research. Project COUNTER operationalizes usage in this way, choosing to track only full-text downloads of journal articles. Project COUNTER, however, does not address a whole range of issues that SSHRC might want to consider. Since journals also perform

functions beyond satisfying information needs of researchers (such as delineating a scholarly community; providing a value system for academic evaluation), an operational definition of use might also encompass social utility or even intrinsic utility (Case, 2002). In contrast to Project COUNTER, the Print Measurement Bureau's (PMB)<sup>7</sup> definition of use takes into account a much broader range of utility. It defines reading as "looking into it [a publication] and reading something, no matter how briefly" (PMB, 2007: v1-2). This operationalization of usage might be too simplistic for SSHRC, providing little means for differentiating between intended and accidental requests to the server.

Ultimately the decision as to what constitutes use of a journal should be addressed by the stakeholder community so that the interests of all parties are accounted for in the evaluation scheme. To initiate this discussion, this chapter begins to unpack the meaning of the metrics identified at the end of Chapter 3 and draws on empirical studies about journal user behaviour to add insight on the validity of conclusions that are drawn based on such metrics.

## **What Can be Measured?**

Based on the information contained in the log file (and in light of the preparation and processing activities identified in Chapter 3), it is possible to identify who has visited the web site, what items they requested, the time at which each request was made, whether the request was successfully fulfilled, and the size of the item delivered by the server<sup>8</sup>. From this empirical material there are a number of other details about the interaction that can be inferred. For instance, the user's trajectory through the site can be

---

<sup>7</sup> The PMB is a Canadian non-profit organization that conducts large-scale readership studies (mostly by survey) of consumer magazines, and more recently, of medical journals <[www.pmb.ca](http://www.pmb.ca)>.

<sup>8</sup> Here I am referring to the common logfile format.

inferred based on the chronology of each request made. The amount of time spent on each page requested may also be inferred, but, as discussed in Chapter 3, doing so may be problematic. This chapter explores what measures can be obtained from a dataset of this scope and what meaning can be extracted as a result.

### **Hits and Files**

The total number of requests made to the server (i.e., each click on a web site) during a given period of time is referred to as *hits*. This metric is easily manipulated because simply clicking on any page from the journal's web site will add to the number of hits. Journal staff could inflate this metric by accessing the web site for normal job-related reasons. Moreover, hits alone provide no information about the quality of the user's interaction with a journal's web site—a request for the journal's home page is not distinguished from requests for full-text articles or table of contents.

*Files* are a subset of hits that result in the request being successfully fulfilled by the server, that is, those requests where a file (any page from the web site, not necessarily a full-text article) was returned by the server and successfully loaded into the browser. Similar to hits, this metric does not provide great insight into the type of interaction that occurred. Both of these metrics should be interpreted only as indicators of traffic reaching the journal's web site.

### **Sessions**

The term *session* is used to describe all of the activities performed by a unique user during a discrete encounter with a web site (see Chapter 3 for details on how sessions are identified and delimited). Sessions are synonymous with *visits*, which is

another term that is sometimes used to describe this metric. Once individual sessions have been identified, analysts can tabulate how many visits to the web site occurred over a period of interest (e.g., total number of visits for the month of January) or they can calculate the average number of visits per day, per week, etc.

Because a session “delineates a number of related information actions, like searching and displaying” (Huntington et al., 2008: 359), the number of sessions conducted is a more meaningful metric than hits or page views. Huntington et al. (2008: 359) note that “in addition to providing simple counts, as a complimentary activity metric to page views, sessions can be used to denote the amount of time that a user spends on a site and the number of pages viewed.” These authors also suggest that the number of sessions can act “as a proxy for user counts” (Huntington et al., 2008: 359).

Although a session delineates a number of related information actions, the number of sessions is really a measure of how many times at least one page of the web site has been accessed by a client browser within a specific period of time. It says something about a site’s visibility in the sense that there is evidence that traffic going to the site. As an indicator of the type of use, however, it is less telling. This metric alone does not give any information on what pages were looked at (i.e., the quality of the interaction), and gives no indication of the extent to which the visitor obtained useful information. Perhaps by filtering for sessions that contain a predefined algorithm of requests it could be used as a measure of ‘legitimate’ or ‘research’ use, if that is the type of interaction that the stakeholder community values.

## Pages Viewed

The log file contains the exact URL of the various items (web pages) requested by the client browser, therefore it is possible to determine which particular pages were viewed and also how many pages were viewed for a given period of time or per visitor. Nicholas et al. (2006) refer to the number of items viewed per online session as *site penetration*, and they suggest that it is a powerful metric. However, these authors also acknowledge that “users who penetrate the site more may be doing so because they cannot find exactly what they want or need and users who view just a few items and leave might do so because they view exactly what they look for and then leave the site” (Nicholas et al., 2006: 1363). In research, finding nothing is not necessarily a bad thing. It may establish that information (relevant literature) is not there, which often results in new research being undertaken to fill the information gap.

The number of single-page visits and the single-page visit rate can also be calculated based on the items viewed. Again, alone these metrics are not exceptionally enlightening. If the single page requested is the journal homepage, then it is probably less likely to be an indicator of research use than if the single page requested was a full-text article. In fact, it may be common for users to request articles directly from an external search engine (e.g., Google) or from a library database (e.g., EBSCO), bypassing the journal web site as a “discovery vehicle” entirely (Institute for the Future, 2002: 14). (EBSCO takes users to its site, not the journal’s web site.)

The authors of eJUS (the e-Journal User Study) identified six major request types for users of medical and biology journals: “Journal Home Page, Table of Contents, Full Text in HTML version, Full Text in PDF format, Search, and selected online

Hyperlinking features (including hyperlinks to cited articles, to articles in press, and to field-specific resources/databases)” (Institute for the Future, 2002: 4). The typology of request types was used to identify patterns of use for e-journals. The three most common usage patterns identified by eJUSt researchers were:

1. Journal homepage → Table of contents → Full text HTML → Full text PDF

2. PubMed → Full text HTML → Full text PDF

3. Journal homepage → Search → Full text HTML → Full text PDF (Institute for the Future, 2002: 9).

Although by and large requests were for full-text articles, these authors noted that the frequency of requests for selected hyperlinking features (e.g., to cited articles, field-specific resources) was still significant, and the importance should not be underestimated. Evaluating these patterns of use led the authors to conclude that “the final goal of most web visits is a PDF version of an article” (Institute for the Future, 2002: 12). This behaviour is reflected in Project COUNTER’s guidelines for collecting usage data; reporting journals are required to show only two metrics: full-text downloads (HTML, PDF, and Total) and turnaways<sup>9</sup> by month. Although this metric is a clear indication of journal use, used alone it is limiting in the type of use that is observed. Given the potential of online journals to provide a range of services beyond article retrieval (see Chapter 2), this metric alone is only part of the picture.

---

<sup>9</sup> Turnaways are defined as requests for full-text articles that were not fulfilled because the site license user limit was exceeded.



## **Duration**

The server does not function as a timer, keeping track of the time spent on each page requested; rather, it timestamps each discrete interaction. If each interaction is viewed in sequence, it is possible, ostensibly, to infer the time spent on each page, a metric referred to as *time by page seen*. It is also possible to calculate *session length* and *average visit duration*.

There are a number of caveats to such time-based metrics because they do not include time spent reading content after leaving the site with a downloaded PDF, and time spent on a page does not equate with time spent reading. Moreover, it is tempting to assume that time spent is proportional to success using the site. This is not necessarily the case, however, because, as with pages viewed, users may in fact spend more time on a site when they are having trouble using it/finding what they need (Nicholas et al., 2006). Johnson, Bellman, and Lohse (2003) found that increased proficiency with a site (due to multiple visits) actually resulted in less time spent per session.

## **Return Visits**

It has been suggested that, in theory, repeated use of a web site “constitutes conscious and directed use,” and that “the number of times someone returns to a site to search...tells us something about site loyalty and satisfaction” (Nicholas et al., 2006: 1358). Indeed, it seems logical to expect that academics who subscribe to a journal “would naturally develop a repeat behaviour in order to fulfil their current awareness needs” (Nicholas et al., 2006: 1358). In practice, however, this is not always the case. In a study of the use of the Blackwell Synergy and Emerald Insight digital libraries, Nicholas et al. (2006) found that 63% to 69% of users made only one visit during the study period

(2 months and 1 year, respectively). This finding is particularly important to a concern with “regular readers.”

Although, as Nicholas et al. (2006: 1358) point out, “it is not clear what would constitute a natural frequency [of use] for a journal site,” several factors have been identified that may influence whether use is repeated. One factor that may be influential is publication frequency. This is because the addition of new content to a web site is known to cause spikes in usage, and for journals, each new issue results in new content becoming available (Nicholas et al., 2007). If a journal publishes quarterly, it might be reasonable to expect 4 visits per year from a loyal reader. More frequently published serials could be expected to have as many visits per individual as there are new issues or far more since doing research would take a person back quite frequently.

Some speculate that user demographics, such as discipline, might have an influence on the frequency with which return visits occur, citing “the more pressing current awareness needs of scientists” as a driving factor for repeated use (Nicholas et al., 2006: 1363). This, of course, presumes that a journal is used for that purpose. In some fields, such as physics, this is not the case.

Because online journals can provide scholars with RSS and email alerts that are tailored to specific information needs, it is not necessary to visit a journal web site regularly to maintain the level of awareness that used to be achieved by receiving and perusing the print issue of the journal. (Hence, RSS recipients should be regarded as a useful complementary measure.) At the same time, scholars may take advantage of these alerting services for more journals than they would normally subscribe to (Meho and Tibbo, 2003), so although they are less likely to visit a particular site often, they are

likely to use a broader selection of journals than in the print era. This behaviour may contribute to what is known as the “long tail.” Long tail theory suggests that more use comes from random one-offs than from concentrated, regular use (Anderson, 2004). This behaviour is particularly evident in online contexts where “massive digital choice” has led to what Nicholas et al. (2006: 1358) refer to as “information promiscuity.” As such, although it is desirable for a journal to have a high proportion of users coming back to the web site often, indicating loyalty and desirability, it is not necessarily something that can be expected given the vast information landscape for scholars to traverse.

Pragmatically speaking, if repeat use is of interest, it may be more meaningful to report this metric as *returning rate*—the proportion of total users that visited the site more than once within a given period of time—than as a count of the number of returning visitors. Such a rate would allow journals with varied user population sizes to be compared, and a minimum standard could be established for evaluation.

Other variables impinge on repeat usage and hence the notion of “regular readers.” Given the relationship of frequency of visits to frequency of content updates, comparison between journals with varying frequencies of content updates introduces inaccuracies. Frequency of visits spurred by content updates may introduce an incentive for journals to move away from issue-based publishing in order to generate regular readers by updating content more regularly. Similarly, the inverse relationship between RSS feeds and frequency of visits may introduce a disincentive to use this valuable research service.

## **Other Issues**

### **Journals as a Delivery Vehicle**

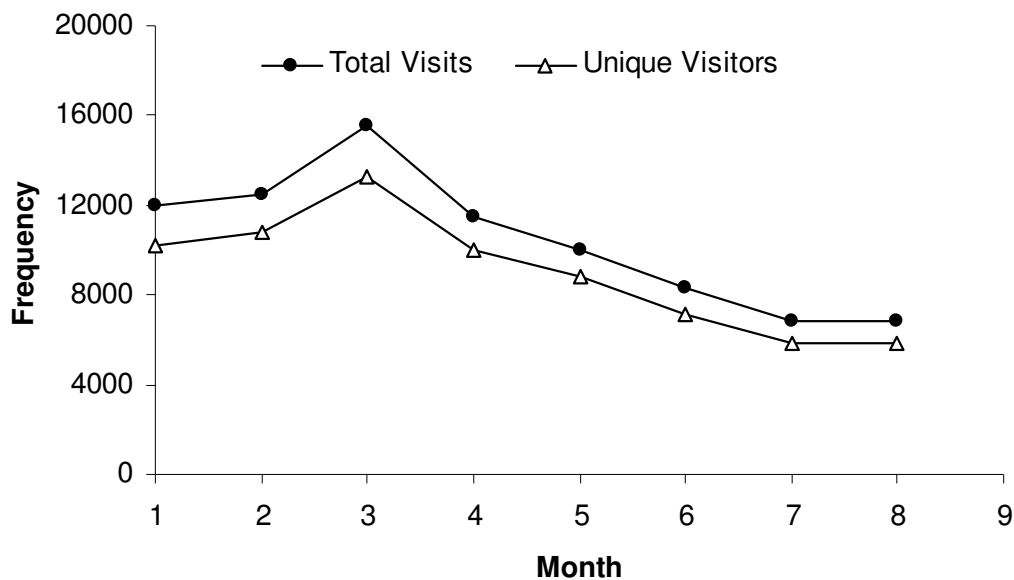
A journal web site functions as an archive for content, but so far all of the methods discussed for measuring usage assume that the journal web site also functions as the sole “discovery and delivery vehicle” for its content (Institute for the Future, 2002). However, this is not always the case. If a journal has licensed its content to an aggregator such as ProQuest, then the content is also hosted by the aggregator. If an article is requested through the aggregator rather than the journal web site, the request will be received and fulfilled by the aggregator’s servers, and its use will not be recorded by the journal and it will not be included in a usage report.

It is important for journals to license their content to aggregators so that it has a chance of being seen by scholars outside of the journal’s primary markets and indeed, sometimes within the journal’s primary market. Thus, it is not advisable to create incentive for journals to restrict access to its content to its own servers simply to ensure that they have enough activity to report for the funding competition. It would be advisable, however, to prompt journals to obtain usage data from the aggregators to whom they have licensed their content. Most aggregators will be able to provide some basic usage information since many of them are COUNTER-compliant.

### **Reporting Period**

Another issue that needs to be addressed when reporting usage metrics is the period for which the metrics were collected. This variable is important because there is evidence that journal usage fluctuates temporally. Monthly data from the *Canadian Journal of Communication* illustrates this phenomenon (Figure 2). Usage of this journal

is greatest between January and April, and is at its highest during March. (Data were not available for September-December.) This pattern of use reflects the academic cycle of semesters, where students (a large portion of this journal’s user base) are likely preparing major research projects for the end of the semester. During the summer months when there are fewer students and academics working, the numbers drop by almost half. If the *CJC* were to report usage from August alone, it would create a very different picture than if it reported usage from March.



**Figure 2. Frequency of total monthly visits and unique visitors to the *Canadian Journal of Communication Online* between January (month 1) and August 2007. (Data source: phpMyVisites)**

Researchers from the eJUSt project analyzed the traffic patterns of 14 life science and medical journals with the aim of identifying discernable cycles of use relating to the journal’s publication frequency. They found that “there was a clear and significant

weekly cycle in terms of traffic (number of sessions) regardless of the length of issue circulation period. Monday and Tuesday had the highest traffic and then traffic slowed down until Friday” (Institute for the Future, 2002: 2). Both of these examples illustrate the importance of acknowledging the period for which data are reported when interpreting web usage metrics.

The SSH journals in Canada are a heterogeneous group, which makes comparing their use difficult on a number of levels. They represent a range of disciplines, from psychology to cultural studies, and these user groups have different information needs and behaviours. As well, each journal is at a different point in its adoption of online publishing. Those journals that have an established online presence will have had longer to tweak their journal web site so that users can get the most out of it. The audience for these journals will be more comfortable accessing it online and will likely have it integrated into their information seeking routines, which can take time—some estimate up to three years for this to occur (Luther, 2001: n.p.). Thus, the usage results for the reporting period where journals are newly establishing an online presence may be expected to be poor.

Generally speaking, evaluating the use of online journals is no simple task. Individually the metrics discussed here represent very limited bits of information about journal usage. But reported together as a group they may be more telling. The amount of traffic that reaches the site is probably best viewed as an indicator of visibility. The difference between the amount of traffic reaching the site and the number of downloaded articles might be a better indicator of the demand for its content. Similarly, the difference between those who penetrate deeply into the site and those that penetrate deeply and

subsequently download articles might also be an indicator of demand for the content, although, penetration can also be an indicator of usability. Duration is probably best viewed as an indicator of usability as well. Measures of repeat use are likely good indicators of loyalty, but should not be given too much weight because of what we know about the likelihood of people to return to the same journal.

## **Conclusions and Recommendations**

Because this study was undertaken in the context of SSHRC's attempt to measure usage and define eligibility for financial assistance based on that usage, the conclusions of this report are cast with respect to SSHRC's actions.

The evolving nature of online journals makes it clear that a holistic approach to evaluating readership is needed. A sound approach to evaluating readership must begin with defining three elements: 1) an approach to measurement, 2) the methods used to obtain data, and 3) the types of metrics used (with their advantages and limitations).

In terms of the overall approach to evaluating journals, SSHRC might consider providing incentive for journals to focus on meeting audience needs. One way of shifting the focus to audience needs may be to measure improvement or sustainability over time. This might entail journals producing monthly usage reports throughout the duration of the funding period so that cycles of use become visible and so that these cycles can be compared over time and triangulated with variables like content updates, changes to web site design, etc. This individualized approach to measurement would reduce the difficulty in comparing numbers from a heterogeneous group of journals and would provide incentive for journals to learn more about their markets in the process of providing evidence of demand.

This report has revealed the great variety of options available for collecting and processing web usage data. The choice of data source (e.g., web server log files, page tagging) and analytics software (e.g., Webalizer, phpMyvisites) can have a significant



impact on the metrics that serve as indicators of demand. SSHRC can manage the potential heterogeneity of reporting methods by choosing a standard web analytics package for journals to use. The two examples of software used in this report were selected as a convenience sample (i.e., the data available were generated using these methods). An alternative program that was not discussed in the report, but which has since emerged as a leader in web analytics is Google Analytics (see <http://www.google.com/analytics/>). This program is free, provides comprehensive documentation and online support, and is very simple to install and use. Most importantly, the program provides excellent analytical tools, including all of the key metrics discussed in this report (visits, unique visitors, page views, pages/visit, one-page visit rate [‘bounce rate’], average time spent on site, % new visits). For this reason, it is becoming widely adopted, and because it is widely adopted and comes from a highly trustworthy source, it is becoming (or at least has the potential to become) a standard benchmark for measurement. Some might say that Google Analytics is even more effective as a circulation audit than any "circulation audit" business in the magazine/advertising world.

Further to this, SSHRC must specify what they consider acceptable configuration settings for the chosen software so that all usage data are processed in the same way. Some of these settings will depend on how usage events are defined—what they can or cannot be (requests for table of contents, abstract, full-text HTML/PDF, search tool, other site features, etc.), so it is important for the stakeholder community to decide how broadly or narrowly usage should be defined.

Once a standard analytics package is chosen, the choice of which metrics to report will easily follow. The biggest challenge for SSHRC is to determine how to interpret whatever the chosen metrics are. It will be important for SSHRC to consider context when assessing each applicant journal, since the *Canadian Journal of Netherlandic Studies* will most probably have a smaller audience of potential users than the *Canadian Journal of Law and Society*. In addition to specifying which metrics to report, it is imperative for SSHRC to specify an interval of time for reporting (e.g., month, semester, year, etc.) so that journals are reporting metrics in comparable units.

Implementing such standards will support journals in their preparation to apply for funding under the current guidelines and will increase the fairness of the competition.

## Reference List

- Anderson, C. (October 2004). The long tail. *Wired*.
- Barrett, B. L. (July 2008). *Webalizer Quick Help*. Retrieved February 16, 2009, from: [http://www.webalizer.org/webalizer\\_help.html](http://www.webalizer.org/webalizer_help.html)
- Bittman, M., Lynch, L., and Pauls, N. (2007). *Best practices guide to scholarly journal publishing*. Vancouver, BC: Canadian Association of Learned Journals.
- Boismenu, G., and Beaudry, G. (2004). Scholarly publishing and public service. *Canadian Journal of Communication*, 29, 343-58.
- Case, D. O. (2002). *Looking for information: A survey of research on information seeking, needs, and behavior*. London: Academic Press.
- Client-server. (n.d.). Retrieved July 25, 2008, from Wikipedia: <http://en.wikipedia.org/wiki/Client-server>
- COUNTER Codes of Practice (2008a). *Release 3 of the COUNTER Code of Practice for Journals and Databases: Appendix D: Guidelines for Implementation*. Retrieved February 13, 2009, from: [http://www.projectcounter.org/code\\_practice.html](http://www.projectcounter.org/code_practice.html)
- COUNTER Codes of Practice (2008b). *Release 3 of the COUNTER Code of Practice for Journals and Databases: The Code of Practice Release 3 August 2008*. Retrieved February 13, 2009, from: [http://www.projectcounter.org/code\\_practice.html](http://www.projectcounter.org/code_practice.html)
- Dainow, B. (April 7, 2005). *3rd party cookies are dead*. <http://www.webanalyticsassociation.org/en/art/?2> Accessed March 3, 2007.
- Ellis, D. (1989). A behavioural approach to information retrieval system design. *Journal of Documentation*, 45, 171-211.
- Huntington, P., Nicholas, D., and Jamali, H. R. (2008). Website usage metrics: A re-assessment of session data. *Information Processing and Management*, 44, 358-372.
- Http cookie. (n.d.). Retrieved February 25, 2009, from Wikipedia: [http://en.wikipedia.org/wiki/HTTP\\_cookie](http://en.wikipedia.org/wiki/HTTP_cookie)
- Institute for the Future. (2002). *E-Journal User Study: Report of Web Log Data Mining*. Retrieved February 24, 2009, from: <http://ejust.stanford.edu/logdata.pdf>

- Johnson, E. J., Bellman, S., and Lohse, J. (2003). Cognitive lock-in and the power law of practice. *Journal of Marketing*, 67(2), 62-75.
- Kaushik (2006). (<http://www.kaushik.net/avinash/2006/12/the-great-web-data-capture-debate-web-logs-or-javascript-tags.html>)
- Lorimer, R. and Lindsay, A. (2004). Canadian scholarly journals at a technological crossroads. *Canadian Journal of Communication*, 29, 253-275.
- Lorimer, R., Lynch, L., and Provençal, J. (2006). *Augmenting print: Planning for online journal publishing by social sciences and humanities journals in Canada*. Vancouver, BC: Canadian Association of Learned Journals.
- Lorimer, R. and Maxwell, J. (2007). Canadian social science and humanities online journal publishing, the Synergies project, and the creation and representation of knowledge. *Publishing Research Quarterly*, 23, 175-93.
- Luther, J. (2001). White paper on electronic journal usage statistics. *Journal of Electronic Publishing*, 6(3). Accessed March 4, 2007. <http://www.press.umich.edu/jep/06-03/luther.html>
- Mackenzie Owen, J. (2007). *The scientific article in the age of digitization*. The Netherlands: Springer.
- Meho, L. I. and Tibbo, H. R. (2003). Modeling the information-seeking behaviour of social scientists: Ellis' study revisited. *Journal of the American Society for Information Science and Technology*, 54, 570-587.
- Nicholas, D., Huntington, P., Jamali, H. R., and Dobrowolski, T. (2007). Characterising and evaluating information seeking behaviour in a digital environment: Spotlight on the 'bouncer'. *Information Processing and Management*, 43, 1085-1102.
- Nicholas, D., Huntington, P., Jamali, H. R., and Watkinson, A. (2006). The information seeking behaviour of the users of digital scholarly journals. *Information Processing and Management*, 42, 1345-1365.
- PMB Print Measurement Bureau. (2007). *PMB Medical Media Study: Questionnaire*. Retrieved February 24, 2009, from the PMB web site: <http://www.pmb.ca/>
- SSHRC. (2007a, July 7). *Aid to Open-Access Research Journals*. Accessed August 21, 2007, from the SSHRC web site: [http://www.sshrc.ca/web/apply/program\\_descriptions/open\\_access\\_journals\\_e.asp](http://www.sshrc.ca/web/apply/program_descriptions/open_access_journals_e.asp)
- SSHRC. (2007b, June 6). *Aid to Open Access Journals Program: Questions for Community Consultation*. Accessed August 21, 2007, from the SSHRC web site:

<https://eservices.nserc.ca/sshrsurvey/TakeSurvey.aspx?SurveyID=16KL8n8&PreviousActualPageNumber=5&PreviousDisplayPageNumber=3&PreviousQuestionNumber=3&ActualPageNumber=6&DisplayPageNumber=4&ResponseID=521&QuestionNumber=3&Edit=No&DisplayHeader=&SP=,3,4>,

SSHRC. (2008, April 22). *Aid to Scholarly Journals*. Accessed April 23, 2008, from the SSHRC web site:  
[http://www.sshrc.ca/web/apply/program\\_descriptions/scholarly\\_journals\\_e.asp](http://www.sshrc.ca/web/apply/program_descriptions/scholarly_journals_e.asp)

Spiliopoulou, M., Mobasher, B., Berendt, B., & Nakagawa, M. (2003). A framework for the evaluation of session reconstruction heuristics in web-usage analysis. *INFORMS Journal on Computing*, 15, 171-90.

Spiliopoulou, M. and Pohle, C. (2001). Data mining for measures and improving the success of web sites. *Data Mining and Knowledge Discovery*, 5, 85-114.

Tenopir, C., King, D. W., Boyce, P., Grayson, M., Zhang, Y., & Ebuon, M. (2003). Patterns of journal use by scientists through three evolutionary phases. *D-Lib Magazine*, 9(5). (no page numbers)

W3C. (July 1995). Logging control in W3C httpd. Retrieved July 25, 2008, from the W3C web site: <http://www.w3.org/Daemon/User/Config/Logging.html>

Web analytics. (n.d.). Retrieved April 15, 2008, from Wikipedia:  
[http://en.wikipedia.org/wiki/Web\\_analytics](http://en.wikipedia.org/wiki/Web_analytics)