

Sampling Strategy

Stephen Wattam

January 11, 2011

1 Introduction

The purpose of this document is to describe the approach to sampling used in the document attrition study.

2 Assumptions

The following assumptions are made of the technical portion of this system:

- The connection to the internet used is representative of a regular home internet user
- HTTP requests are made to websites in the same manner as that of a regular home user (with common client software)
- When a site is inaccessible or changed to the sampling machine, it is changed globally
- Downloading data at different times of the day does not affect page content
- The incidence of concurrent connections to other items on the list is of no consequence to each site's served data.

3 Sampling Strategy

The samples taken are a longitudinal sample of all websites within the data set (once weekly), using the same computer. This emulates a user accessing their bookmarks once weekly until they cease to exist.

Each sample is identifiable both by sequence number and by the exact time at which it was downloaded.

3.1 Data Stored

insert listing from implementation