

Document Attrition2

Stephen Wattam

January 26, 2011

1 Introduction

Bookmarking has become a core part of behaviour when browsing the web — from the earliest browsers, users have chosen to retain direct links to content they find useful or think they might revisit. This has developed into globally-accessible services such as del.icio.us, where users may share bookmarks and sync them between computers.

The purpose of bookmarking, however, is frequently thwarted by the nature of content online — some pages change purpose often, websites may be purchased or repurposed, or the site may go down entirely. Indications from other studies imply strongly that the problem of document attrition online is a significant one, to the degree that even small delays between bookmarking and recall may lead to significant document attrition.

The tendency of bookmarks to point at single pages may add to this effect: a page is likely to retain its TLD and/or subdomains as they form part of the branding and access method, but its internal page structure is often subject to change.

2 Research Questions

1. Is document attrition a significant issue on bookmarking services' purpose?
2. Assuming the bookmark sample is representative of the web, how much skew does this effect upon web-as-corpus studies?

3 Bookmarking Services

Some bookmarking services now download and archive a copy of the web page. Explore these.

1. Pinboard (<http://pinboard.in/>)

2. Historius (<http://historio.us/>)
3. Diigo (<http://www.diigo.com/>)

4 Data Gathering

Starting with a random sample of bookmarks from an online bookmarking service, data must be longitudinally sampled over a period of many years, with each snapshot being stored for later access.

Many aspects of the data are important, however, the content remains king. The ease with which text may be processed (compared to screenshots or pure HTML) indicates that initial efforts should concentrate on:

- Page HTML code, from which content may be stripped
- Page content within that HTML (ignore boilerplate and comments)

This data is relatively easy and quick to sample, and places little recurring load on the server.