

LiWA: Living Web Archives

Timestamping Pages on on the Web

Pierre Senellart



UNIVERSITÉ
PARIS-SUD 11

May 2008

HTTP Timestamping

Two timestamping mechanism in HTTP: **entity tags** and **modification dates**. Potentially provided with all requests:

Etag: "497bef-1fcb-47f20645"

Last-Modified: Tue, 01 Apr 2008 09:54:13 GMT

Etag: unique identifier for the provided document, should change if the document changes; can be used in requests with If-Match and If-None-Match.

Last-Modified: last modification time; can be used in requests with If-Modified-Since and If-Unmodified-Since.

- Information generally provided and very reliable for static content (often includes media files in CMS).
- Information hardly ever provided (or with dummy *now* date) for dynamic content, CMS, etc.

HTTP Cache and Proxy Control Mechanisms

Two additional (redundant) items of **freshness information**, for caches and proxies:

```
Cache-Control: max-age=60, private
```

```
Expires: Tue, 01 Apr 2008 13:25:55 GMT
```

`max-age`: maximum time in second a document remains fresh.

`Expires`: time when a document stops being fresh.

- Often provided...
- ...but with 0 or very low expiration delay.
- \Rightarrow does not give any interesting information.

Timestamps in HTML Web page Content

- **Very frequent** in CMS:
 - either as a **global** timestamps (*Last modified:*);
 - or on **individual** items: news stories, blogs, etc. (is the global timestamp the max of the individual ones?);
 - possibly also in meta-data on the Web page: comments, Dublin Core <meta> tags.
- Quite easy to identify and extract from the Web page (keywords, date recognition).
- Informal: sometimes partial (no time indication), often without timezone.
- Might not always be trustworthy.

Additional semantic timestamps

Files of other types than HTML may have **semantic** timestamping mechanism:

PDF, Office suite documents, etc.: both **creation** and **modification** date available in metadata. Quite reliable.

RSS feeds: reliable **semantic** timestamps.

Images, Sounds: **EXIF** (or similar) metadata. Not always reliable, and the capture date of a picture may have nothing in common with its publication date.

Additional semantic timestamps

Files of other types than HTML may have **semantic** timestamping mechanism:

PDF, Office suite documents, etc.: both **creation** and **modification** date available in metadata. Quite reliable.

RSS feeds: reliable **semantic** timestamps.

Images, Sounds: **EXIF** (or similar) metadata. Not always reliable, and the capture date of a picture may have nothing in common with its publication date.

Semantic external content used for dating a HTML Web page:

- Possibility of mapping a **RSS feed** to a Web page content to date individual items.
- **Sitemaps** provided by the Web site owner. Allows for providing both timestamps and change rate indications (*hourly*, *monthly*...), but these functionalities are not often used. A few CMS produce all of this: **ideal case!**

Estimating Freshness of a Page

- 1 Check HTTP timestamp.
- 2 Check content timestamp.
- 3 Compare a hash of the page with a stored hash.
- 4 Non-significant differences (ads, fortunes, request timestamp):
 - only hash text content, or “useful” text content;
 - compare distribution of n -grams (shingling);
 - or even compute edit distance with previous version.

Adapting strategy to each different archived website?

Conclusion

Estimating freshness and timestamp of a Web page: **possible in most cases**. May require to retrieve the entire page multiple times (not always possible to do conditional GETs).