Trimming the fat: THA Big Bad Cleanup of my WordPress

THA Big Bad Cleanup of my WordPress: from importing tweets & statuses to posting in Markdown.

About two *(or more)* months ago I decided to pull myself together from the shards[^1]:

- import ~~all~~ my important Tweets *(thank God altogether there's less than 800)*
- import important Facebook updates & posted photos
- dig up all the saved version of all my previous sites and get most of the posts from there
- remove unnecessary plugins
- clean up leftover shortcodes, old markup, etc. from old posts

Importing

Twitter

Someone thought about exporting tweets in a usable manner: you'll receive a csv version of all your tweets when you go to Settings[^2] and click **Request your archive**. All you need is WP Ultimate CSV Importer[^3] and you can even import the retweet and reply metadata from the csv file; although you might need a bit of a magic to replace the links with the expanded ones instead of the shortlinks.

Facebook

On the other hand, Facebook is terrible. The export you receive is an HTML file, the entries aren't in a block but separated with `<div class="meta"` elements. If you've ever used unusual clients to post to Facebook with (like Friendcaster), some entry will look like this:

```
Sunday, 18 September 2011 at 15:09 UTC+01
Péter Molnár shared a link.
(the comment I added to the link)
```

and no actual link at all. This is a manual process to import and it hurts.

freedom.io

I gave freedom.io[^4], a tool to import from silos to blogs a go, but it simply did not work for me: error message, no import. :(

All my previous blogs

I made PHP4 code run with 5.4 PHP-FPM. I digged archive.org[^5] for long lost data. I encountered how bad HTML was when there was no IE6 yet.

Do this as soon as possible; it's getting more and more troublesome with every day.

Still: the earliest I could do was 2002. It seems the Hungarian internet before that was gone; vanished with most of the free hosting providers. I was also rookie enough not to make incremental backups or version controlling that time and kept overwriting the page with updates, so the initial 3 years seem to be lost forever. (I'll try to dig up some 15 year old CDs but I don't believe in this kind of miracle.)

Lessons learnt: keep incremental backups and keep lots of backups of these backups.

Analytics

I decided to drop the Jetpack provided WordPress Analytics and go ~~with Piwik~~[^6]~~: a self-hosted analytics tool and will probably hammer my MySQL backend, but I give it a go.~~ with Slim Stat Analytics[^7]. The WordPress Analytics is not providing enough information (like time spent on the page), and I'm trying to opt out from Google services, so this is the way I have to go for now.

Markdown & cleanup

Markdown never really won me until very recently, when I finally realized how much easier is to read & edit it and how much more futureproof it is compared to HTML. I probably needed to see my HTML4, XHTML, HTML5 content mixed together for a bit of a shock; and try Ghost[^8], where the editor is a win over anything. *The system itself is way too restricted for me though.*

Fortunately I've never used really tricky markup in my posts, but using regular HTML to Markdown[^9] did not work well on the content conversion.

On the frontend, the original PHP Markdown failed: no definition lists, no proper Github Flavour, and they were not handling the images the way I wanted. So I went with Markdown Extra[^10] dialect and Parsedown[^11] on the frontend display.

Update: forget my solution, just use Pandoc[^12].

I'm only leaving the PHP here for historical reasons.

I had roughly 3 HTML tables in 650 posts, that I converted by hand; for all the others, I wrote my mini-converter.

Disclaimer: I'm well aware this could be done much better. I'd also like to warn you to make backups of your database and not to run this script on production.

```php
<?php

function html2markdown ( $content ) {

    if (empty($content))
        return false;

    $hash = sha1($content);
    if ( $cached = wp_cache_get ( $hash, __CLASS__ .
__FUNCTION__ ) )
        return $cached;

    $content = preg_replace('#\s(id|class|style|rel|data|
content)="[^"]+"#', '', $content);
    /*
     * Credits to @gnarf
     * https://stackoverflow.com/questions/3026096/remove-all-attributes-from-an-html-tag
     *
        /              # Start Pattern
         <             # Match '<' at beginning of tags
         (             # Start Capture Group $1 - Tag Name
          [a-z]         # Match 'a' through 'z'
          [a-z0-9]*     # Match 'a' through 'z' or '0' through '9' zero or more times
         )             # End Capture Group
         [^>]*?        # Match anything other than '>', Zero or More times, not-greedy (wont eat the /)
         (\/?)         # Capture Group $2 - '/' if it is there
         >             # Match '>'
        /i            # End Pattern - Case Insensitive
     *
     */
    //$content = preg_replace("/<([a-z][a-z0-9]*)[^>]*?(\/?)>/i",'<$1$2>', $content);

    /**
     * replace <pre>, <code>, [code] and [cc]
     */

    if ( strstr( $content, '<pre><code>' )) {
        $s = array ( '<pre><code>', '</code></pre>' );
        $r = array ( "```\n", "\n```" );
        $content = str_replace ( $s, $r, $content );
    }

    if ( strstr( $content, '<pre>' )) {
        $s = array ( '</pre><pre>', '</pre>' );
        $r = array ( "```\n", "\n```" );
        $content = str_replace ( $s, $r, $content );
    }

    if ( strstr( $content, '</code>' )) {
        $s = array ( '<code>', '</code>' );
        $r = array ( "```\n", "\n```" );
        $content = str_replace ( $s, $r, $content );
    }
```

```
    // straigtforward formatting: html to markdown
    $s = array ( '<tt>', '</tt>', '<bold>', '</bold>', '<strong>', '</strong>', '<em>', '</em>', '<i>', '</i>' );
    $r = array ( '`', '`', '**', '**', '**', '**', '*', '*', '*', '*' );
    $content = str_replace ( $s, $r, $content );

    $s = array ( '<p>','</p>', '<br />', '<br>', '<h1>', '</h1>', '<h2>', '</h2>','<h3>', '</h3>','<h4>', '</h4>','<h5>', '</h5>','<h6>', '</h6>', '<blockquote>', '</blockquote>' );
    $r = array ( "\n", "\n", "\n", "\n", '#', '', '## ', '', '### ', '', '#### ', '', '##### ', '', '###### ', '', '> ', '' );
    $content = str_replace ( $s, $r, $content );

    preg_match_all('/<ul>(.*?)< \/ul>/s', $content, $uls);
    if ( !empty ( $uls[0] ) ) {
        foreach ( $uls[0] as $to_replace ) {
            $to_clean = preg_replace ( '/\t<li>/', '- ', $to_replace );
            $s = array ( '</li>', '</ul><ul>', '</ul>', '<li>' );
            $r = array ( '', '', '', '- ' );
            $to_clean = str_replace ( $s, $r, $to_clean );
            $content = str_replace ( $to_replace, $to_clean, $content );
        }
    }

    preg_match_all('/<ol>(.*?)< \/ol>/s', $content, $ols);
    if ( !empty ( $ols[0] ) ) {
        foreach ( $ols[0] as $to_replace ) {
            $to_clean = $to_replace;
            preg_match_all('/<li>(.*?)< \/li>/s', $to_clean, $lis);
            foreach ( $lis[0] as $id=>$lis_replace ) {
                    $liline = $lis_replace;
                    $lis_replace = preg_replace ( '/\t</li><li>/', $id+1 . '. ', $lis_replace );
                    $lis_replace = preg_replace ( '/</
li><li>/', $id+1 . '. ', $lis_replace );
                    $to_clean = str_replace ( $liline ,
$lis_replace, $to_clean );
            }

            $content = str_replace ( $to_replace, $to_clean,
$content );
        }
    }

    $s = array ( '<ol>', '</ol>', '</li>' );
    $r = array ( '', '', '' );
    $content = str_replace ( $s, $r, $content );

    preg_match_all('/<dl>(.*?)< \/dl>/s', $content, $dl);
    if ( !empty ( $dl[0] ) ) {
        foreach ( $dl[0] as $to_replace ) {
            $to_clean = $to_replace;
            preg_match_all('/<dt>(.*?)< \/dt>/s', $to_clean,
$dts);
            preg_match_all('/<dd>(.*?)< \/dd>/s', $to_clean,
$dds);

            foreach ( $dts[0] as $id=>$dt ) {
                    $o_dt = $dt;
                    $o_dd = $dds[0][$id];

                    $dt =  str_replace ( array('<dt>', '</
dt>' ), array( "" , "\n" ), $dt );

            }
        }
    }

    $c = str_get_html ( $content );
    if (!$c)
        return $content;

    // find links
    foreach($c->find('a') as $a) {
        $out = $href = $title = $txt = '';
        $href = $a->href;
        $title = $a->title;
        $txt = $a->innertext;

        if ( !empty( $txt ) && !empty ( $href ) ) {
            if (!empty($title))
                $out = '['. $txt .' '.$title.']('. $href .')';
            else
                $out = '['. $txt .']('. $href .')';
            $content = str_replace ( $a->outertext, $out,
$content );
        }
    }

    // clean up images:
    foreach($c->find('img') as $img) {
        $src = $alt = $title = $cl = $out = false;

        $src = $img->src;
        $alt = $img->alt;
        $title = $img->title;

        if ( empty($alt) && !empty($title) ) $alt = $title;
        if ( empty($alt) ) $alt = $src;

        $img = '!['.$alt.']('. $src;
        if ( !empty($title) ) $img .= ' '. $title;
        $img .= ')';

        $content = str_replace ( $img->outertext, $img,
$content );
    }

    // fix potential hashtag issues
    $content = preg_replace ( '/^#/mi', '\#', $content );

    wp_cache_set ( $hash, $content, __CLASS__ . __FUNCTION__,
static::expire );
```

```
    return $content;
}
```

So far, so good. Some plugins are not playing that well with the Markup-only text I'm now saving into, but I'll probably just post patches to them for this.

Links

1. https://petermolnar.net/indieweb-decentralize-web-centralizing-ourselves/
2. https://twitter.com/settings/account
3. https://wordpress.org/plugins/wp-ultimate-csv-importer/
4. http://www.freedom.io/
5. http://archive.org
6. https://petermolnar.net/piwik.org
7. https://wordpress.org/plugins/wp-slimstat/
8. https://ghost.org
9. https://github.com/nickcernis/html-to-markdown
10. https://github.com/tanakahisateru/js-markdown-extra
11. http://www.parsedown.org/demo?extra=1
12. http://pandoc.org/scripting.html

Created by Peter Molnar <mail@petermolnar.net>, published at 2014-07-18 22:50 UTC, last modified at 2021-10-31 15:57 UTC , to canonical URL https://petermolnar.net/article/wordpress-cleanup-markdown-import-twitter-import-facebook/ , licensed under CC-BY-4.0 .