How to audit the entire site for a specific term

Sometimes we need to search the whole site for a specific term to find all instances of it. This might be to replace the term with something else, like if a department name changes, or remove references to it entirely.

While you might be tempted to use Find and Replace to do this automatically, it's important to consider the context in which the term is used, as you might inadvertently change something you didn't mean to.

> **Pain**
>
> This is a very tedious process and can be very time-consuming, so consider first whether it's really required. Weigh the costs and benefits. Weren't you going to retire to that cabin by the lake? It's your granddaughter's piano recital tonight and you promised you'd be there...

Step-by-step guide

Plan your attack

Identify all versions of the term you want to search for, including possible variations in capitalisation.

Decide what needs to be done with the terms - what needs to be changed and what can stay the same?

> Don't remove references to the term in news articles, as they would have been accurate at the time of publication.

Inform your colleagues that you are going outside and you may be some time.

Searching the CMS

1. Create a new spreadsheet for your audit. I'd suggest the following columns:
   - URL in CMS
   - Term (including variations)
   - Line
   - Who is responsible for the page
   - Action taken
2. Use the find content in pages tool to search for your term. You will need to search all the top-level folders separately.
3. The tool will show a list of the pages that have the term, the term in context and the line number. You will need to enter this into the spreadsheet manually.
4. Continue searching all the folders and fill in the spreadsheet as you go. Remember to search for different variants of the phrase, for example RDSO and R.D.S.O.

Searching the W: drive

1. Ask a dev to:
   1. grep the filesystem for each of your terms (this will take hours and should be done overnight). For example:

      ```
      # on www0
      cd /www/vhosts/bath/
      ggrep -ri 'rdso' * > /tmp/output1.txt
      ggrep -ri 'research development support' * > /tmp/output2.txt
      ```

   2. *There should be something we can do automatically here to make the output less awful, but it would need some requirements*
   3. Collect the output files and send them to you
2. Get a few great big .txt files.
3. Create a new Excel file, or a new tab in your existing spreadsheet.
4. On the Data tab, select "From Text" and then your .txt file.
5. Import the file as a spreadsheet, using : as the separating character to divide the cells.
6. Tidy up your spreadsheet.
   - Search for URLs containing "old", "test" or "webarchive" and remove those rows.
   - The W: drive audit will include every instance of the term, not just every page, so you may want to delete duplicates of pages.
7. Add additional columns for:
   - Who is responsible for the page
   - Action taken.

Archive all the rubbish

You've probably unearthed a lot of outdated pages and documents that are no longer in use but are still taking up space. Archive them.

Related articles

- Customise a GeckoChat 'live' chat channel
- Recover from a git disaster
- Getting started with Grunt
- How to use Flow to manage an editorial calendar
- How to audit the entire site for a specific term