Diagnosing Web Site Architecture Issues
SMX London 2009

Richard Baxter
In house - UK SEO Manager at Cheapflights.co.uk

Blogger and SEO Consultant at SEOgadget.co.uk

Likes: Battlestar, Vanilla ice cream and Lib:Tech snowboards
Diagnosing indexing problems using Google queries
Yahoo Site Explorer
Canonicalization (Checklist)
Duplicate content surprise
Important checks you should do with your web server
Tracking URLs and indexed development servers
Content churn in recruitment and product driven websites
“Dynamic” navigation with internal 301 redirects

DIAGNOSIS: ARCHITECTURE
Use Google queries

- Start with site:yourdomain.tld

- Check specific URL groups and if ok, remove

- Continue to break down by subfolder and intitle – look at every “theme” on your site

- Inspiration – draw your site architecture and use as a URL guide. Think content groups.
What are we looking out for?

**Rail Jobs:** Railway Job Vacancies UK
Rail jobs. Search for jobs in railways for project management, engineer, safety, quality, contracts manager and more. Apply today.
www.edenbrown.com/rail-jobs/ - 65k - Cached - Similar pages

*In order to show you the most relevant results, we have omitted some entries very similar to the 1 already displayed.*
*If you like, you can repeat the search with the omitted results included.*

You end up with:

**Rail Jobs:** Railway Job Vacancies UK
Rail jobs. Search for jobs in railways for project management, engineer, safety, quality, contracts manager and more. Apply today.
www.edenbrown.com/rail-jobs/ - 65k - Cached - Similar pages

**Rail Jobs:** Railway Job Vacancies UK
Rail jobs. Search for jobs in railways for project management, engineer, safety, quality, contracts manager and more. Apply today.
www.edenbrown.com/rail-jobs - Similar pages
Use Google Queries

* Sense check the number of indexed pages
* Drill down using inurl: intitle: and filetype:
* Check for legacy filetypes and old documents
Download the first 1000 indexed URLs

Crawl list with Xenu – you’ll be surprised!

Use “maximum level” set to 0
**WWW’s**
- [http://www.seogadget.co.uk](http://www.seogadget.co.uk)
- [http://seogadget.co.uk](http://seogadget.co.uk)

**Trailing slashes (on or off, just not both)**
- [http://seogadget.co.uk/seo-jobs](http://seogadget.co.uk/seo-jobs)
- [http://seogadget.co.uk/seo-jobs/](http://seogadget.co.uk/seo-jobs/)

**Lower or title case redirects**
- [http://seogadget.co.uk/SEO-Jobs/](http://seogadget.co.uk/SEO-Jobs/)

**How to do 301 redirects in Apache .htaccess, IIS, PHP, ASP and ColdFusion**
Case study - Pagination

Product page: Cameras

- http://www.outdoorpros.com/Cat/Cameras/1241/List
☆ Product page: Cameras

http://www.outdoorpros.com/Cat/Cameras/1241/List?first_answer=13
Fix duplicate content

* Make sure your internal links are consistent
  - Sitemaps, content and navigational links all lower case
  - 301 redirect your www – decide which way and stick to it
  - 301 your trailing slash in or out and stick to it

* Hide analytics and tracking with onclick

```html
<a onclick="location.href='http://seogadget.co.uk/?utm_source=dupe-source'; return false;"
   href="http://seogadget.co.uk">Seogadget home page</a>
```

* Noindex, follow your paginated pages

```html
<title>Wordpress | SEOGadget | Category Archive</title>
<meta http-equiv="content-type" content="text/html; charset=UTF-8" />
<meta name="author" content="Richard Baxter" />
<meta name="robots" content="noindex,follow" />
```
Add this `<link>` tag to specify your preferred version:

```html
<link rel="canonical" href="http://www.example.com/product.php?item=swedish-fish"/>
```

Helps to solve this problem but not the answer!

We know Dubai is quite a long way from the UK but you can always try and get a cheap flight out there, perhaps you still have some holiday to take? Check out British Airways or Cheap Flights for some last minute deals. Once you are there you can always hire a car as the event is providing a Park and Ride service for tennis fans heading to the Dubai Tennis Stadium this year. So it couldn't be easier really. Go on, you only live once.
**Web server header checklist**


<table>
<thead>
<tr>
<th>Started</th>
<th>Time</th>
<th>Sent</th>
<th>Received</th>
<th>Method</th>
<th>Result</th>
<th>Type</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>00:00:00.901</td>
<td>0.362</td>
<td>834</td>
<td>174</td>
<td>GET</td>
<td>200</td>
<td>text/html</td>
<td><a href="http://domainnamewire.com/">http://domainnamewire.com/</a></td>
</tr>
<tr>
<td>00:00:01.496</td>
<td>0.012</td>
<td>821</td>
<td>(439)</td>
<td>GET</td>
<td>(Cache)</td>
<td>text/css</td>
<td><a href="http://domainnamewire.com/wp...kmarks/social_bookmarks.css">http://domainnamewire.com/wp...kmarks/social_bookmarks.css</a></td>
</tr>
<tr>
<td>00:00:01.537</td>
<td>0.013</td>
<td>785</td>
<td>(4969)</td>
<td>GET</td>
<td>(Cache)</td>
<td>application/javascript</td>
<td><a href="http://domainnamewire.com/wp-i...ludes/js/tw-sack.js?ver=1.6.1">http://domainnamewire.com/wp-i...ludes/js/tw-sack.js?ver=1.6.1</a></td>
</tr>
</tbody>
</table>

- Test internal pages give the correct response
  - Particularly your 404 error page...

- Introducing an awesome combination of SEO tools....
<table>
<thead>
<tr>
<th>#</th>
<th>Source URL</th>
<th>Linking Root Domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>searchenginewatch.com/</td>
<td>03.15.00</td>
</tr>
<tr>
<td>2</td>
<td>searchenginewatch.com/repsd/articles.php/216_</td>
<td>280.00</td>
</tr>
<tr>
<td>3</td>
<td>searchenginewatch.com/repsd/article.php/217_</td>
<td>312.00</td>
</tr>
<tr>
<td>4</td>
<td>searchenginewatch.com/webmasters/article.php/218_</td>
<td>394.00</td>
</tr>
<tr>
<td>5</td>
<td>searchenginewatch.com/showPage.html?page=0332_</td>
<td>297.00</td>
</tr>
<tr>
<td>6</td>
<td>searchenginewatch.com/showPage.html?page=0319_</td>
<td>288.00</td>
</tr>
<tr>
<td>7</td>
<td>searchenginewatch.com/showPage.html?page=0320_</td>
<td>287.00</td>
</tr>
<tr>
<td>8</td>
<td>searchenginewatch.com/3071991</td>
<td>280.00</td>
</tr>
<tr>
<td>9</td>
<td>searchenginewatch.com/awards/2002-winners.html</td>
<td>280.00</td>
</tr>
<tr>
<td>10</td>
<td>searchenginewatch.com/showPage.html?page=0318_</td>
<td>273.00</td>
</tr>
<tr>
<td>11</td>
<td>searchenginewatch.com/limits/</td>
<td>271.60</td>
</tr>
<tr>
<td>12</td>
<td>searchenginewatch.com/awards/</td>
<td>271.60</td>
</tr>
<tr>
<td>13</td>
<td>searchenginewatch.com/awards/article.php/330_</td>
<td>255.00</td>
</tr>
<tr>
<td>14</td>
<td>searchenginewatch.com/webmasters/</td>
<td>240.00</td>
</tr>
</tbody>
</table>

9th Most linked to page is a 404

280 Linking Roots, Linkscape May 2009
Other web server stuff:

Is your development server indexed?

Don’t leak your staging URLs into Google!

Edwards, Kenny & Bray LLP - Page | Edwards, Kenny & Bray LLP
PROFESSIONAL DEVELOPMENT. Professional excellence is one of the core values on
which our firm was founded. Our commitment to professional excellence not ...

Health and sustainable development | Article | CABE
15 May 2006 ... Anna Coote, head of Engaging Patients and the Public at the Healthcare
Commission, explores ways to strengthen the relationship between the ...

Millenium Development Goals | Make Poverty History
Submitted by Holly on November 11, 2007 - 7:00pm. Today’s always a weird day for me. A
hard day for some reason - pardon the melodrama. ...
For sites that sell clicks or leads to advertisers

http://www.comparestoreprices.co.uk/visit.asp?v=Argos&u=https://www.|ksrv%60%60%60/click-1081665-5590799%ACurl=http%3A%2F%2Fwww%2Eargos%2Eco%2Eu%2Fwe...%2Fstoreld%3D10001%26catalogId%3D651%26langId%3D%2D1%26go%2Ex%3D4%26go%2Ey%3D11%26searchTerms%3D568%2F6917&m=Sony&p=PlayStation%203%20Console%2040GB

User-agent: *
Disallow: /includes/
User-agent: aipbot
Disallow: /
User-agent: BecomeBot
Disallow: /
User-agent: psbot
Disallow: /
User-agent: Roverbot
Disallow: /
Content churn

* Handle your product / jobs pages nicely

Pages respond with a 200 server header, no content on page, <title> and meta description are gone.
This job could not be found

Search for another job
Keep your pages alive, but make the user experience valuable

The vacancy’s been placed, but we’ve got plenty of others!

You just have to solve the internal linking problem
“Dynamic” navigation

* Navigational links point to:
  – /content/dynamic.asp?nodeid=237

* Which 301 redirects to:
  – /project-management-pmo-jobs/

* AND has a session ID in the URL

* Avoid internal 301 redirects
Thank you

Twitter: @richardbaxter
Email: richard@seogadget.co.uk
Download: http://seogadget.co.uk/presentations