

XML-sitemap optimizations for large sites: a case study.

Track: Clients & Industry Experiences

Liopold Novelli¹ Mag. Jeremy Chinquist¹

¹drunomics G.m.b.H.

Oct 5, 2021 / DrupalCon Europe

drunomics

Table of Contents

- 1 Introduction
- 2 Research & Current State-of-the-Art
- 3 The Solution
 - Results
- 4 Conclusion
 - Summary
 - References

Outline

- 1 Introduction
- 2 Research & Current State-of-the-Art
- 3 The Solution
 - Results
- 4 Conclusion
 - Summary
 - References

Background

Over the past years, multiple clients have requested better SEO and sitemap indexing.

Die Wirtschaftsverlag, a publishing house in Vienna, Austria, has a range of websites, including Handwerk+Bau. The case study will address Handwerk+Bau.

Background

The current state of Handwerk+Bau:

- A single sitemap.xml file with all URLs, available at `https://www.handwerkundbau.at/sitemap.xml`
- The file was evaluated by search engines on a daily basis, but we were uncertain how effective it was with 5k urls.
- Some URLs were not processed for days.

The Goals

How can the sitemap be optimized for a website with large amounts of content?

We found the key here:

By providing the last modification timestamp, you enable search engine crawlers to retrieve only a subset of the Sitemaps in the index i.e. a crawler may only retrieve Sitemaps that were modified since a certain date. This incremental Sitemap fetching mechanism allows for the rapid discovery of new URLs on very large sites...

sitemaps.org 2016

The Goals (2)

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
<!-- Generated by the Simple XML Sitemap Drupal module: https://drupal.org/project/simple_sitemap. -->
▼<sitemapindex xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  ▼<sitemap>
    <loc>_____general/sitemap.xml</loc>
    <lastmod>2021-10-02T09:06:30+00:00</lastmod>
  </sitemap>
  ▼<sitemap>
    <loc>_____monthly/2021-10-1/sitemap.xml</loc>
    <lastmod>2021-10-02T09:06:30+00:00</lastmod>
  </sitemap>
  ▼<sitemap>
    <loc>_____monthly/2021-09-1/sitemap.xml</loc>
    <lastmod>2021-10-02T09:06:30+00:00</lastmod>
  </sitemap>
  ▼<sitemap>
    <loc>_____monthly/2021-08-1/sitemap.xml</loc>
    <lastmod>2021-10-02T09:06:30+00:00</lastmod>
  </sitemap>
```

Outline

- 1 Introduction
- 2 Research & Current State-of-the-Art
- 3 The Solution
 - Results
- 4 Conclusion
 - Summary
 - References

Research

- Schema.org
 - sitemaps.org 2016
- Resources concerning index & crawl rates and past experience.
 - Central 2021

Research

- Resources concerning what sitemap structure is and when one requires a sitemap.
- Related (case) studies
 - Pilgrim 2007
 - Lee et al. 2009
 - Manickam 2014

Goals

This produced the following short list of possible changes:

- Create multiple *sitemap* files to list content. Keep these sitemap files small and group logically.
- Create multiple sitemap *index* files.
- Each sitemap index file must use lastmod of when the individual sitemap was modified.

In short, to date we achieved all but the last goal. Analysis of this change will come in a future presentation.

Requirements to Implement the Solution

Thus a (Drupal-specific) solution was required to:

- Create individual sitemaps with similar content (i.e. by node type)
- Group the sitemaps by some logic
- Add the individual sitemaps to auto-generated sitemap index files. The solution must automatically break files with 50000 items into sub-files.

Current Drupal State-of-the-Art

- Simple XML Sitemap (Contrib Module) Tymchuk and Ginalski 2021
 - We already had a simpler version of an extension module that provided custom sitemap indexes.
- XML Sitemap (Contrib Module) Reid 2019
 - Lacks customization versatility
- Custom module

Outline

- 1 Introduction
- 2 Research & Current State-of-the-Art
- 3 The Solution**
 - Results
- 4 Conclusion
 - Summary
 - References

The Implemented Solution

- Extend Simple XML Sitemap (Contrib Module) Tymchuk and Ginalski 2021
- Add a new (contrib) module called Simple Sitemap Extensions Mueller, Novelli, and Chinquist 2021 introducing ...
 - Custom sitemap index
 - Plugin types:
 - SitemapType - variant
 - SitemapGenerator
 - UrlGenerator

The Implemented Solution

- ... sitemap variants via
`/admin/config/search/simplesitemap/variants`
- ... node-type settings per sitemap variant. E.g.
`/admin/structure/types/manage/article`
- The Simple XML Sitemap module gives us the opportunity to sort sitemap links via `hook_simple_sitemap_links_alter`. This is not included in Simple Sitemap Extensions. A custom Wirtschaftsverlag module was implemented for sorting.

The Implemented Solution

After creating the module and installing, we can define sitemap variants.

The screenshot shows the Drupal 8 administration interface for the 'Simple XML Sitemap' module. The breadcrumb trail is 'Startseite > Administration > Konfiguration > Suche und Navigation > Simple XML Sitemap'. The page title is 'Simple XML Sitemap'. Below the title, there is a message: 'If you would like to say thanks and support the development of this module, a donation will be much appreciated.' The main content area is titled 'SITEMAP VARIANTS' and explains that a sitemap variant is a specific instance of a certain type. It lists 'VARIANTEN' (variants) with their bundle settings, custom links, and corresponding sitemap instances. The variants listed are: 'handbau_general' (path: /sachliche_sachliche_entity | Handwerk und Bau - General), 'handbau_sachliche' (path: /sachliche_sachliche_entity | Handwerk und Bau - Sachliche_sachliche), and 'handbau_positiv_sachliche' (path: /sachliche_sachliche_entity | Handwerk und Bau - Sachliche_sachliche). Below this, there is a section for 'Examples' showing how to define a variant and its label. The 'Available sitemap types' section lists: 'default_hreflang' (default hreflang sitemap type), 'indexed_entity' (allows linking images within paragraphs and altering the entity query), 'dynamic_sitemap_type' (dynamic sitemap type), 'sitemap_index' (sitemap index type), and 'monthly_dynamic_type' (dynamic sitemap to show articles by month).

The Implemented Solution

Then tell Drupal to include / exclude node types from the variants.
Channel pages appear in the general sitemap.

The screenshot shows the Drupal configuration page for 'Inhaltstyp Channel page bearbeiten'. The main content area is titled 'Sitemap variants' and contains two variant configurations:

- HANDWERK UND BAU - GENERAL**
 - Do not index entities of type Channel page in variant Handwerk und Bau - General/
 - Index entities of type Channel page in variant Handwerk und Bau - General/
 - PRIORITY**
 - 0.5
 - Die Priorität, die Entitäten dieser Art für Suchmaschinen-Bots haben.
 - ÄNDERUNGSFREQUENZ**
 - Nicht festgelegt -
 - Die Häufigkeit, mit der sich Entitäten dieses Typs ändern. Suchmaschinen-Bots können dies als Hinweis darauf nehmen, wie oft sie indiziert werden sollen.
 - BILDER MITTEINBEZIEHEN**
 - Nein
 - Legt fest, ob Bilder, auf die von Entitäten dieses Typs verwiesen wird, in die Sitemap aufgenommen werden sollen.
- HANDWERK UND BAU - MONTHLY ARTICLES**
 - Do not index entities of type Channel page in variant Handwerk und Bau - Monthly articles/
 - Index entities of type Channel page in variant Handwerk und Bau - Monthly articles/

The Implemented Solution

Articles are further distributed accross monthly sitemaps.

The screenshot shows the Drupal administration interface for 'HANDWERK UND BAU'. The top navigation bar includes 'Startseite', 'Verwalten', 'Verknüpfungen', and a search bar. Below the navigation bar, there are several menu items: 'Inhalt', 'Struktur', 'Design', 'Erweitern', 'Konfiguration', 'People', 'Berichte', and 'Hilfe'. The main content area is divided into two columns. The left column contains settings for the 'Eingabeformular', 'Veröffentlichungseinstellungen', 'Spracheinstellungen', 'Anzeigeinstellungen', 'Menüinstellungen', 'Simple XML Sitemap', and 'Zeitplaner'. The right column is titled 'Sitemap variants' and contains two sections: 'HANDWERK UND BAU - GENERAL' and 'HANDWERK UND BAU - MONTHLY ARTICLES'. In the 'HANDWERK UND BAU - GENERAL' section, the 'Do not index entities of type Artikel in variant Handwerk und Bau - General' option is selected, indicated by a red arrow. In the 'HANDWERK UND BAU - MONTHLY ARTICLES' section, the 'Index entities of type Artikel in variant Handwerk und Bau - Monthly articles' option is selected, also indicated by a red arrow. Below these sections, there are settings for 'PRIORITÄT' (0.5) and 'ÄNDERUNGSFREQUENZ' (Nicht festgelegt). At the bottom of the page, there is a link for 'HANDWERK UND BAU - SITEMAP INDEX'.

Results

In the previous 2 months we have seen that the individual sitemaps are consistently and frequently crawled.

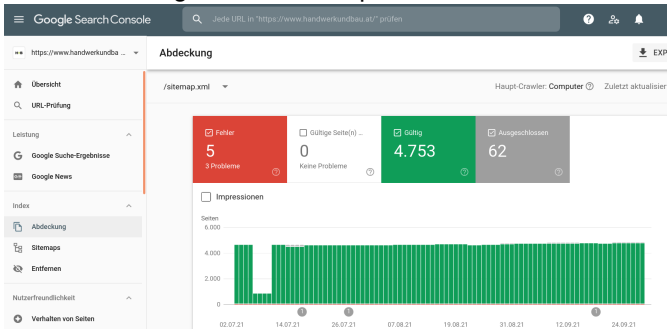
Sitemaps > /sitemap.xml

SITEMAP ÖFFNEN 

Sitemap	Zuletzt gelesen	Status	Gefundene URLs
/sub/hand-bau_general/sitemap.xml	28.09.2021	Erfolgreich	167
/sub/hand-bau_monthly/2013-01-1/sitemap.xml	27.09.2021	Erfolgreich	15
/sub/hand-bau_monthly/2013-02-1/sitemap.xml	26.09.2021	Erfolgreich	20
/sub/hand-bau_monthly/2013-03-1/sitemap.xml	23.09.2021	Erfolgreich	15
/sub/hand-bau_monthly/2013-04-1/sitemap.xml	22.09.2021	Erfolgreich	20
/sub/hand-bau_monthly/2013-05-1/sitemap.xml	22.09.2021	Erfolgreich	19
/sub/hand-bau_monthly/2013-06-1/sitemap.xml	22.09.2021	Erfolgreich	23
/sub/hand-bau_monthly/2013-07-1/sitemap.xml	27.09.2021	Erfolgreich	15
/sub/hand-bau_monthly/2013-08-1/sitemap.xml	25.09.2021	Erfolgreich	20

Results

The number of indexed pages is 98.61% and this has been consistent since the change to the sitemap structure:



Outline

- 1 Introduction
- 2 Research & Current State-of-the-Art
- 3 The Solution
 - Results
- 4 **Conclusion**
 - Summary
 - References

Conclusion




- Google consistently and frequently crawls the sitemaps
- The index rate is acceptable at > 98%

- Roadmap
 - Fix the lastmod date of a sitemap index file entry in the main sitemap index file and complete the use-case analysis.
 - Move sitemap link sorting logic into the Simple Sitemap Extensions module.
 - Make Simple Sitemap Extensions compatible with Simple XML Sitemap 4.x.

References

-  Mueller, Mathias, Liopold Novelli, and Jeremy Chinguist (2021). *Simple Sitemap Extensions*. URL: https://www.drupal.org/project/simple_sitemap_extensions (visited on 09/30/2021).
-  Reid, Dave et. al (2019). *XML sitemap*. URL: <https://www.drupal.org/project/xmlsitemap> (visited on 12/18/2019).
-  Tymchuk, Andrey and Pawel Ginalski (2021). *Simple XML sitemap*. URL: https://www.drupal.org/project/simple_sitemap (visited on 07/03/2021).

References

-  Lee, Hsin-Tsang et al. (June 2009). “IRLbot: Scaling to 6 billion pages and beyond”. In: *ACM Transactions on the Web (TWEB)* 3, pp. 427–436. DOI: [10.1145/1367497.1367556](https://doi.org/10.1145/1367497.1367556).
-  Manickam, Chandran (Jan. 2014). “A Study on Website Quality Evaluation based on Sitemap”. In:
-  Pilgrim, Chris (Jan. 2007). “Trends in sitemap designs - A Taxonomy and survey”. In: *Conferences in Research and Practice in Information Technology Series* 64.