Sitemap là gì?

Sitemap là bản đồ của một website, là tập tin văn bản chứa tất cả URL của một website. Chứa dữ liệu về mỗi URL. Sitemap hướng dẫn các search engine thu thập các thông tin của website một cách hiệu quả hơn và đồng thời cập nhật thay đổi khi website có thay đổi.

Trong file xml sitemap có chứa thông tin url, thời gian update và trọng số Priority

Thông số priority quy định tính quan trọng của url đối với trang web

Khi nào cần Sitemap

Khi website thật sự lớn

Website có 1 kho lưu trữ lớn các trang tách riêng hoặc không liên kết với nhau

Website mới và ít liên kết bên ngoài

Website có sử dụng đa phương tiện được hiển thị trong google tin tức hoặc sử dụng các chú thích tương thích với sơ đồ website khác

Referenece: https://www.sitemaps.org/protocol.html

Web crawlers usually discover pages from links within the site and from other sites. Sitemaps supplement this data to allow crawlers that support Sitemaps to pick up all URLs in the Sitemap and learn about those URLs using the associated metadata.

Tài liệu Sitemap là một XML Schema

Một Sitemap bao gồm:

<urlset></urlset> phải có namespace

Bắt buộc phải có, đóng gói têpk và tham chiểu tới The Current Protocol Standard

<url></url>

Bắt buộc phải có. Là thẻ cha của các thẻ <loc> và là 1 entry cho XML tag

<loc></loc>

Bắt buộc phải có. Chứa url của trang của website. Các url phải bắt đầu bằng các giao thức (http hoặc https) và kết thúc bằng dấu gạch chéo

<lastmod></lastmod>

Tuỳ chọn. Ngày sửa đổi của file. Định dạng YYYY-MM-DD cái này khác với IF-Modified-Since (304) ở Header.

<changefreq></changefreq>

Tuỳ chọn. Tần suất thay đổi trang có các giá trị:”always, hourly, daily, weekly,monthly, yearly, never”. Giá trị always sử dụng cho các tài liệu thay đổi mỗi khi chúngh đc truy cập. Giá trị never được sử dụng để mô tả các URL lưu trữ. Giá trị thẻ chỉ là gợi ý.

<priority></priority>

Tuỳ chọn, Mức độ ưu tiên của Url này liên quan đến url khác trên website. Các giá trị này từ 0.0 đến 1.0. Mức độ ưu tiên là tương đối, nó chỉ được sử dụng để chọn giữa các URL trên trang website đó.

Entity escaping

Mỗi file sitemap phải encoded UTF-8. Mỗi file xml, các dữ liệu bắt buọc phải dử dụng entity escape code:

|  |  |  |
| --- | --- | --- |
| Ampersand | & | &amp; |
| Single Quote | ' | &apos; |
| Double Quote | " | &quot; |
| Greater Than | > | &gt; |
| Less Than | < | &lt; |

Ngoài ra, tất cả các URL (bao gồm URL của Sơ đồ trang web của bạn) phải được URL-escape và được mã hóa để có thể đọc được bởi máy chủ web mà chúng được định vị. Tuy nhiên, nếu bạn đang sử dụng bất kỳ loại tập lệnh, công cụ hoặc tệp nhật ký nào để tạo URL của mình (bất kỳ thứ gì ngoại trừ việc nhập chúng bằng tay), điều này thường được thực hiện cho bạn.

Mỗi một sitemap không được lớn hơn 50MB và hơn 50000 URL. Muốn có nhiểu sitemap thì phải có Sitemap index file, là doc liệt kê danh sách các sitemap cũng ko đc lớp hơn 50MB và 50.000 URL và chứa sitemap của cùng website.

Bao gồm

<sitemapindex></sitemapindex> Bắt buộc, thông tin của tất cả các sitemap trong sitemap index file

<sitemap></sitemap> Bắt buộc, thông tin của từng sitemap

<loc></loc>Bắt buộc, Xác định vị trí của sitemap bằng URL, RSS file, Atom file

<lastmod> Tùy chọn, Xác định thời gian mà sitemap tương ứng đã được sửa đổi.

Khó