Anke Probst

MÖGE DIE MACHT MIT DEN SITEMAPS SEIN!

Teil 1: Crawl- und Indexierungsmanagement

DIE AUTORIN

Anke Probst verantwortet als Senior-SEO-Managerin die SEO-Strategie der XING SE. Sie ist außerdem Teil des Expertenbeirats des Bundesverbands Digitale Wirtschaft (BVDW) e. V. und Co-Autorin des Buches „Der Online Marketing Manager".

Webmaster und SEOs sind sich grundsätzlich einig: Sitemaps sind kein direkter Rankingfaktor. Folglich gibt es wichtigere Baustellen auf der SEO-Roadmap, die positiv auf Rankings und Traffic einzahlen. Ist das nicht etwas zu vorschnell gedacht? Sitemaps haben durchaus eine Berechtigung, aus dem Schatten hervorzutreten und höhere Priorität zu genießen. Im ersten Teil dieses Zweiteilers stellt Anke Probst dar, warum, wie und für welche Webseiten sich der Aufwand eines durchdachten Sitemap-Konzeptes durchaus sehr lohnen kann – mit einem Praxisbeispiel für erfolgreiches Crawl- und Indexierungsmanagement. Der Folgeartikel in der nächsten Ausgabe zeigt dann auf, wie Sie Sitemaps zur Analyse nutzen können, um Optimierungspotenziale zu finden.

Illu: axel2001 / gettyimages

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <url>
    <loc>http://www.example.com/foo.html</loc>
    <lastmod>2018-06-04</lastmod>
  </url>
</urlset>
```

Abb. 1: Beispiel einer einfachen Sitemap mit den erforderlichen Einträgen URL und <lastmod>

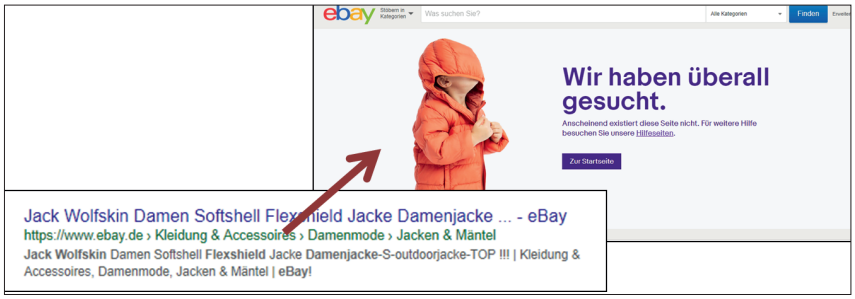


Abb. 2: Veralteter/gelöschter Content im Google-Index ist bei großen Domains keine Seltenheit und führt häufig auf 404-Seiten

TIPP

Achtung: Das Bereitstellen einer Sitemap garantiert nicht, dass alle darin eingereichten URLs und deren Inhalt gecrawlt bzw. indexiert werden oder sogar SEO-Traffic bringen – Suchmaschinen betrachten sie in der Regel als eine Empfehlung bzw. als Hilfestellung, da sie die aufgelisteten URLs eventuell gar nicht oder nur mit viel Zeit und Aufwand crawlen würden. Das heißt im Klartext: Eine saubere Seitenarchitektur ist die absolute Grundlage!

Manchmal hat man den Eindruck, wenn es um SEO geht, werden ein paar grundlegende Dinge vergessen. Es fließt viel Aufwand in Contentproduktion, Optimierung von Ladezeiten und aktuelle Hype-Themen wie Voice Search - aber altbekannte, „langweilige" SEO-Basics wie Sitemaps werden gerne stiefmütterlich behandelt. Das kann ein großer Fehler sein! Gerade bei mittelgroßen bis sehr großen Seiten und/oder komplexen Strukturen ergeben sich viele Fragen, die sich leicht mithilfe von Sitemaps beantworten lassen – und deren Beantwortung die SEO-Strategie häufig stark beeinflusst. Das ist aber längst nicht alles: Mit einem klugen Sitemap-Set-up können nicht nur wertvolle Erkenntnisse, sondern auch kurzfristig zusätzlicher SEO-Traffic gewonnen werden.

Eines sollten Sie nie vergessen: Grundlage für die Indexierung relevanter Inhalte (Text, Medien wie Videos oder Bilder) ist das Crawling. Die stark vereinfachte SEO-Formel lautet: Nur was gecrawlt wird, kann indexiert werden, ranken und Traffic bringen. Das Crawling muss also effizient gesteuert werden – und was ist hierfür besser geeignet als strukturierte Listen von URLs, die den Suchmaschinen direkt zum Crawling und damit auch zur Indexierung zugeführt werden? Bevor Sie sich also den hippen SEO-Themen widmen, denken Sie an die Grundlagen und stellen Sie die folgenden Fragen an Ihren Content:

1. **Wie viele Ihrer Inhalte sind indexiert und sind diese wirklich relevant?**
   Site-Abfragen sind ungenau, die Daten in der Search Console zur Indexabdeckung können oft nur schwer eingeschätzt werden. Wie nähert man sich also dieser Fragestellung?
2. **Könnte es sein, dass ein Teil der relevanten Inhalte gar nicht indexiert ist?**
   Technische Hürden, fehlerhafte Implementierungen von Noindex-Anweisungen, komplexe Seitenstrukturen, falsche Statuscodes und viele weitere Gründe könnten hier vorliegen. Wie aber erkennen Sie die Muster?
3. **Sind Ihre bereitgestellten Inhalte vielleicht kurzlebig oder sollten aus anderen Gründen nur kurzzeitig indexiert sein?**
   Beispiele wären Kleinanzeigen, Events, ständig wechselndes Produktsortiment – Content also, der schnell indexiert werden muss, aber auch schnell wieder veraltet.
4. **Findet frischer Content in tiefen Ebenen statt, z. B. als neues Produkt in einem breit gefächerten Sortiment, als Beitrag in einem Forum oder als relevanter (!) Kommentar zu einem Dauerbrenner-Artikel? Und wird an diesen Stellen vielleicht seltener gecrawlt, sodass die Inhalte erst spät indexiert werden?**
5. **Welche Arten von Content/Medien sind für Suchmaschinen erklärungsbedürftig und benötigen Meta-Informationen, um besser verstanden und bewertet zu werden?**
   Bilder und Videos können noch immer schwer von Suchmaschinen verstanden werden.

Nicht zuletzt helfen Ihnen diese Überlegungen, eine Grundfrage zu beantworten: Was genau braucht Ihre Domain in Bezug auf Crawling, damit relevanter Content indexiert wird und ranken kann? Betrachten Sie es außerdem aus dem Blickwinkel der Suchmaschinennutzer: Wo fallen Ihnen selbst Probleme auf, wenn Sie nach Ihrem Content suchen? Hören Sie auf Ihre User/Kunden: Gibt es eventuell Beschwerden, z. B. dass veralteter Content (ausverkaufte Produkte, längst vergangene Events usw.) rankt, wenn

nach aktuellen Themen gesucht wurde? Dieses Feedback kann enorm wichtig sein, um Probleme bei der Indexierung zu identifizieren, und nicht selten ist das Crawling die Hauptursache.

XML-Sitemaps können diese Fragestellungen gut beantworten bzw. unterstützen. In Kombination mit RSS-/Atom-Feeds sind sie das Mittel der Wahl für effiziente Crawling- und Indexierungssteuerung, wie das folgende Beispiel zeigt.

Ein Beispiel aus der Praxis: xing.com

Die Domain xing.com mit aktuell ca. 25 Millionen indexierten URLs ist mit ihrer komplexen Seitenstruktur sicherlich kein Regelfall, hier lässt sich der positive Effekt von XML-Sitemaps aber gut nachvollziehen. Einige Eckpunkte der Gegebenheiten: Ein paar Millionen URLs waren innerhalb einer flachen und sauber strukturierten Architektur verlinkt. Eigentlich gut durchdacht für effizientes Crawling! Zusätzlich war eine Standard-XML-Sitemap vorhanden - rein formal schien also alles in Ordnung zu sein. Doch immer wieder tauchten vereinzelt Probleme mit der Indexierung auf, u. a. wurde neuer Content innerhalb dieser Struktur nicht schnell genug indexiert bzw. veralteter/gelöschter Content nicht schnell genug deindexiert.

Bei Seiten solcher Größenordnung ist effizientes Crawling an den richtigen Stellen immer eine Herausforderung, und so drängte sich der Verdacht auf, dass die punktuell auftretenden Indexierungsprobleme vielleicht nur die Spitze des Eisbergs sein könnten. Besonders die oben genannten Fragen 1-4 sind für xing.com ungemein wichtig: Man verliert bei dieser Menge an URLs den Überblick über die Anzahl wirklich relevanter, indexierungswürdiger URLs.

Die Fragen zu schnell veraltendem und ständig neuem, aber tief verankertem Content ergeben sich aus der Masse an UGC (User Generated Content): User erwarten, dass ihr Content möglichst schnell indexiert wird, der neue Content liegt jedoch meist auf Klickebene 4-5, die nicht zügig genug gecrawlt wird. Ebenso wird aber erwartet, dass gelöschte Inhalte, insbesondere wenn sie personenbezogen sind, sehr schnell auch nicht mehr über Suchmaschinen auffindbar sind – eine Erwartungshaltung, der man so gut wie möglich zuarbeiten muss! Zur Unterstützung des Crawlings wurde daher ein neues, zur Domain passendes Sitemap-Konzept entworfen – unter genauer Einhaltung der folgenden Basics.

```
<?xml version="1.0" encoding="UTF-8"?>
  <sitemapindex xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <sitemap>
     <loc>http://www.ihrebeispielurl.de/sitemap1.xml.gz</loc>
     <lastmod>2004-10-01T18:23:17+00:00</lastmod>
  </sitemap>
  <sitemap>
     <loc>http://www.ihrebeispielurl.de/sitemap2.xml.gz</loc>
     <lastmod>2005-01-01</lastmod>
  </sitemap>
  </sitemapindex>
```

Abb. 3: Einfaches Beispiel einer Index-Sitemap, die zwei Sitemaps enthält

```
https://www.heine.de/robo
User-agent: *
Sitemap: https://www.heine.de/sitemap.xml
Disallow: /webapp/
Disallow: /JSGlobals
Disallow: /EcondaGlobals
Disallow: /ProductJSON
Disallow: /kleingeschenke/
Disallow: /MobileMiniBasketSolo
#20170628
```

Abb. 4: Die Datei robots.txt von heine.de mit Angabe der Index-Sitemap

Basics und Tipps: Erstellen und Einreichen einer Sitemap

Der Aufbau einer Sitemap ist unter sitemaps.org verständlich dokumentiert und das Protokoll wird von allen gängigen Suchmaschinen unterstützt. Neben dem XML-Protokoll finden sich hier auch die grundlegenden Definitionen für RSS-Feeds und Textdateien, die ebenfalls als Sitemap-Formate nutzbar sind. Ergänzend dazu ein paar wichtige Basics und Tipps:

Suchmaschinenspezifische Anforderungen

Ein Blick in die Hilfestellungen der Suchmaschinen (Google: *bit.ly/2wJTxzH,* Bing: *binged.it/1RHdHz5*) lohnt sich, um spezifische Anforderungen oder Eigenheiten zu berücksichtigen. Beispielsweise gibt Google konkret an, dass die Angabe <priority> nicht mehr unterstützt wird, <lastmod> ist jedoch weiterhin unbedingt erforderlich.

XML/RSS/Atom ...?

XML ist das gängige Format, wenn sich der Content selten ändert. Wenn jedoch häufig neuer Content entsteht oder upgedatet wird (z. B. Blogs) oder man auf eine schnelle Indexierung angewiesen ist, sollte man auf RSS-Feeds/Atom zurückgreifen. Diese erfüllen ebendiesen Zweck erfahrungsgemäß sehr gut! Google empfiehlt daher ausdrücklich beide Formate: *https://bit.ly/2e4zLZ9.*

Inhalt der Sitemaps

Wichtigste Anlaufstelle der Dokumentationen ist die Auflistung der erforderlichen und optionalen XML-Tags und deren Attribute. Beschränken Sie sich nicht auf die Mindestanforderung, sondern stellen Sie so viele sinnvolle (Meta-)Informationen wie möglich bereit! Dies gilt vor allem für Medien-Sitemaps. Sprechen Sie ggf. mit dem Produktmanager, um eine Übersicht aller vorhandenen Informationen zu erhalten, die dann in eine Sitemap aufgenommen werden könnten. Wie bei allen SEO-Maßnahmen gilt: Halten Sie sich mit Keywordstuffing zurück, dies wird Ihnen im Zweifel eher auf die Füße fallen.

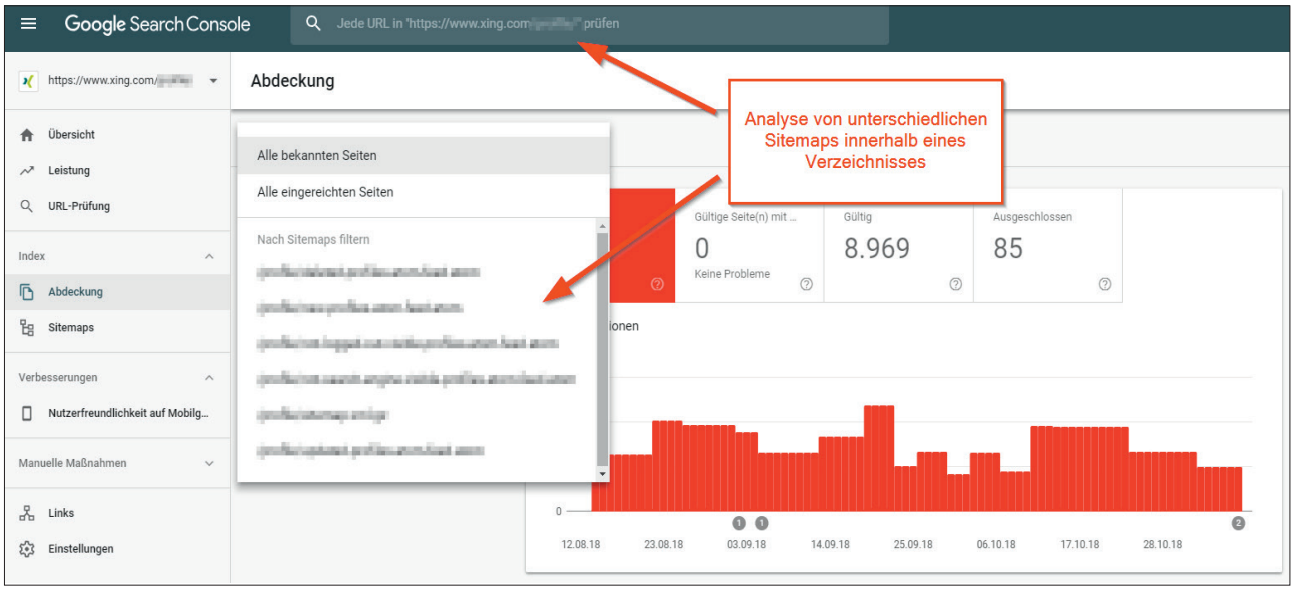


Abb. 5: Beispiel einer Property mit mehreren Sitemaps, die hier spezifische Analysemöglichkeiten bieten

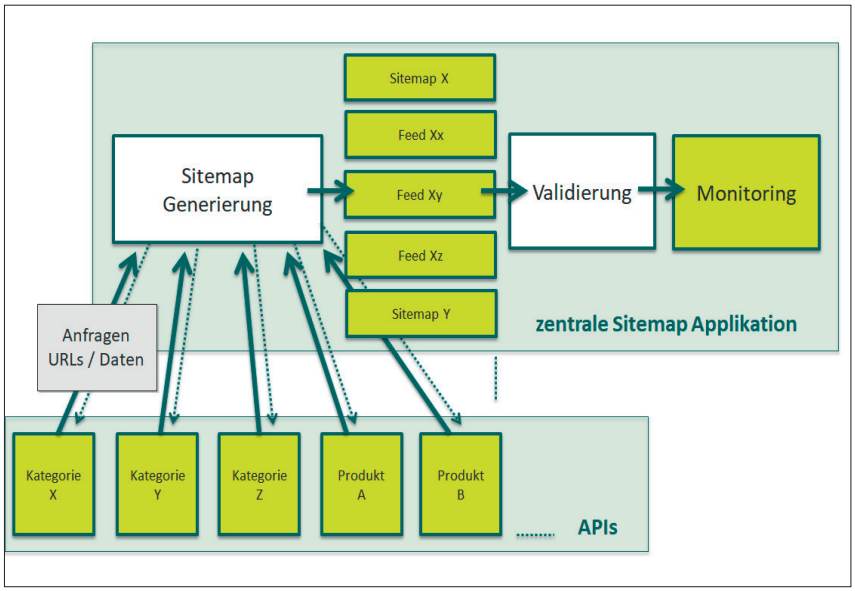


Abb. 6: Stark vereinfachte Darstellung einer Sitemap-Applikation

Sauber arbeiten!

Erstellen Sie saubere, fehlerlose Sitemaps mit validem XML. Stellen Sie sicher, dass nur zur Indexierung freigegebene, kanonische URLs mit dem Statuscode 200 in die Sitemap eingebunden werden – also keine URLs, die doppelten Content erzeugen (Parameter, Canonicals, Session-IDs, Filter-URLs), paginierte Seiten, URLs mit der Meta-Robots-Angabe „noindex", per robots.txt gesperrte URLs, Fehlerseiten (Statuscodes 404/410) oder Weiterleitungen (301/302). Wenn eine Sitemap nicht sauber ist, kann es passieren, dass Suchmaschinen ihr nicht vertrauen und sie ignoriert wird.

Erwähnenswert ist eine Ausnahme für Feeds: Feeds können sich hervorragend zur schnellen Deindexierung eignen. Veralteter Content, der naturgemäß nicht mehr verlinkt wird, wird meist selten gecrawlt und kann für Suchmaschinennutzer ein echtes Ärgernis sein. Nutzen Sie daher Feeds für effizientes Indexmanagement, indem Sie spezielle Feeds mit URLs generieren, die den Statuscode 410/404 oder ein Meta-Tag „noindex" haben. Achten Sie jedoch auf ein genaues Monitoring der Effekte!

Sitemaps für Medien und News

Sie können extra Sitemaps für Medien (Videos, Bilder) generieren oder diese Medien in den normalen Sitemaps zusätzlich bereitstellen. Nutzen Sie jedoch die spezifischen XML-Tags, die Sie in den oben verlinkten Dokumenten nachlesen können. Für News sollten Sie definitiv eine eigene Sitemap verwenden, denn für diese gelten spezifische Anforderungen, die Sie hier nachlesen können: *bit.ly/2qGRyJG*. Voraussetzung ist, dass Ihre Website erfolgreich in Google News aufgenommen wurde.

Sinnvolle Aufteilung von Sitemaps

Eine Sitemap darf nur maximal 50.000 URLs enthalten und nicht komprimiert höchstens 50 MB groß sein. Bei großen Seiten müssen die URLs also auf mehrere Sitemaps aufgeteilt werden – eine prima Gelegenheit, sich um eine sinnvolle Aufteilung Gedanken zu machen. Behalten Sie im Hinterkopf, dass Sie nicht nur die Suchmaschine mit URLs füttern möchten, sondern Sitemaps auch zur Beantwortung Ihrer Fragestellungen nutzen wollen. Eine Aufteilung sollte Ihnen also bestenfalls die Auswertung vereinfachen, z. B. in der Googles Search Console. Folgende Szenarien (aber auch Mischformen) haben sich bewährt:

- Orientierung an der Seitenstruktur. Einzelne Sitemaps für Kategorien bis zur zweiten Ebene können Ihnen später bei der Auswertung in der Google Search Console dank besserer Übersichtlichkeit zugutekommen. In einem Online-Shop für Schuhe mit der Kategorisierung /damenschuhe/sandalen würde die entsprechende Sitemap also lediglich die URLs und Meta-Informationen jener Produkte enthalten, die in dieser Kategorie angesiedelt sind. Diese Kategorie-Sitemaps können Sie in das entsprechende Verzeichnis hochladen, also z. B. unter *www.beispielshop.de/damenschuhe/sandalen/sandalen-sitemap.xml*.
- Sitemaps nach Seitentyp. Je nach Sitemap sind Kategorieseiten, Produkte, Landingpages, Forenbeiträge etc. enthalten.
- Sitemap nach Qualität: Wenn Sie Content in unterschiedlicher Qualität haben und aufgrund dessen Indexierungsprobleme fürchten, sollten Sie die Inhalte nach Qualitätskriterien splitten. Denkbar sind z. B. Sitemaps für Inhalte mit/ohne Bilder oder abhängig von der Contentmenge.
- Wie oben erwähnt: unterschiedliche Feeds zu Deindexierung oder zur schnellen Indexierung neuer Inhalte, z. B. einzelne Feeds für abgelaufene Anzeigen (Statuscode 410), neue Forenkommentare, bestehende, aber aktualisierte Inhalte.

Fassen Sie alle Einzel-Sitemaps der Kategorien in einer Index-Sitemap zusammen, wie im Sitemap-Protokoll dokumentiert (Abbildung 3).

Aktualisierung

Es versteht sich von selbst: Um von Sitemaps profitieren zu können, müssen diese stets auf dem aktuellen Stand sein. Ansonsten drohen Fehler oder die gewünschten Effekte von De-/Indexierung werden nicht erreicht. Aktualisieren Sie die Sitemaps also regelmäßig, mindestens einmal täglich – sofern sich mindestens täglich etwas am Content ändert. Eine News-Sitemap sollte sich sofort aktualisieren, wenn neue Artikel erscheinen oder bestehende Artikel aktualisiert werden und eine neue URL generiert wird.

Für Feeds empfiehlt Google, den Dienst PubSubHubbub zu nutzen, um Updates schnell an Google zu übermitteln (Quelle: *https://bit.ly/2e4zLZ9*, PubSubHubbub: *https://bit.ly/2zMfBet*).

Typ	Inhalt	Anforderung	Grund
XML-Sitemap	» Produkt-URL » Bilder-URL	» Produkt existiert länger als X Tage » Statuscode 200 » Meta-Robots-Index » Kanonische URL	Muss nicht regelmäßig gecrawlt werden, gehört aber zum indexierbaren Inventar
Feed	» Produkt-URL	» Neues, vor < X Tagen im Shop aufgenommenes Produkt » Statuscode 200 » Meta-Robots-Index » Kanonische URL	Neue Produkte müssen möglichst schnell indexiert werden
Feed	» Produkt-URL	» Vor < X Tagen aktualisiert » Statuscode 200 » Meta-Robots-Index » Kanonische URL	Aktualisierter Inhalt (Beschreibung des Produkts, Userbewertung) muss möglichst schnell neu gecrawlt werden
Feed	» Produkt-URL	» Vor < X Tagen gelöschtes Produkt » Statuscode 404/410 » Kanonische URL	Gelöschte Produkte sollen möglichst schnell deindexiert werden
… beliebig erweiterbar	…	…	…

Tabelle 1

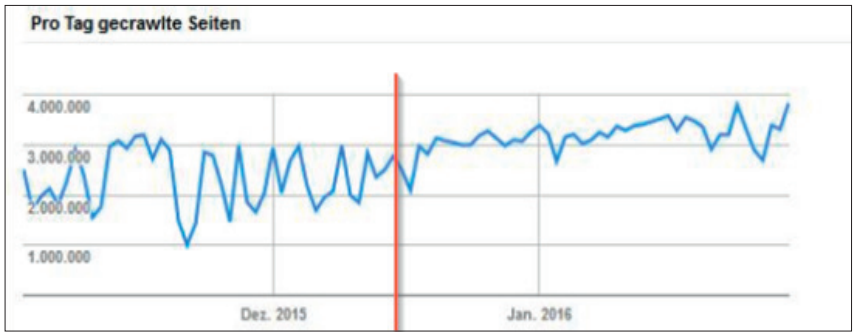


Abb. 7: Stabilisierung der pro Tag gecrawlten Seiten. Markierung: Launch der Applikation

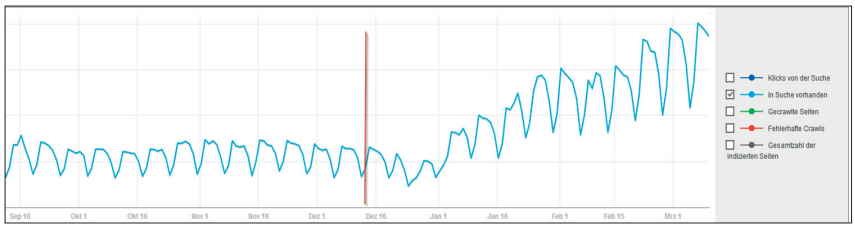
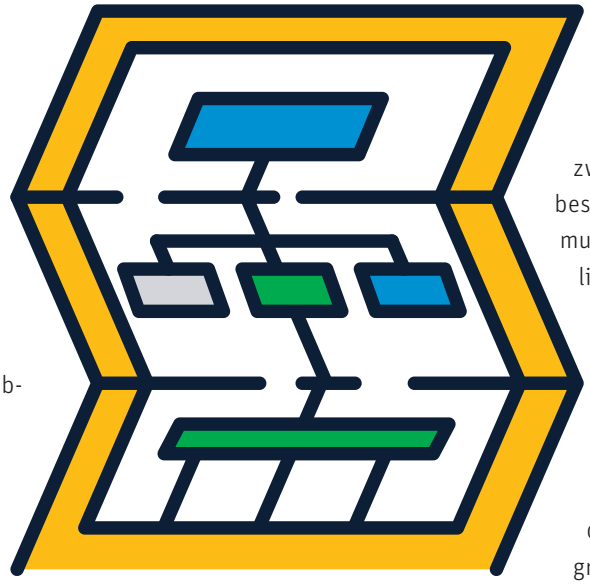


Abb. 8: Anstieg der Impressions bei Bing. Markierung: Launch der Applikation

Zurück zum Praxisbeispiel: das passende Sitemap-Konzept

Nach Sichtung aller relevanter Contentformate auf xing.com wurde klar: Die bestehende Sitemap genügt den Anforderungen der Inhalte nicht. Es werden viele unterschiedliche Inhalte veröffentlicht, die aus diversen Gründen schnell de-/indexiert werden müssen, die in unregelmäßigen Abständen aktualisiert werden, komplett wegfallen oder an anderer Stelle sogar nur eine begrenzte Lebensdauer



haben. Das Ziel muss immer sein, diese Inhalte so gut wie möglich für Crawling und Indexierung zu handhaben - es ist also ein gut definierter Anforderungskatalog nötig, der alle Inhalte und deren Eigenschaften berücksichtigt.

Ein denkbarer schlichter Anforderungskatalog für einen Webshop könnte beispielsweise wie in Tabelle 1 zu sehen ist aussehen. Dies bildet bei XING die Grundlage für das Konzept einer zentralen Sitemap-Applikation, die spezifische Sitemaps generiert, täglich aktualisiert und Suchmaschinen über die Änderungen informiert (Abbildung 6).

Die Applikation aggregiert URLs und alle Sitemap-relevanten Daten von anderen Applikationen der Plattform über APIs. So werden Sitemaps und Atom-Feeds für News, Bilder, unterschiedliche Seitentypen aus verschiedenen Produkten/Kategorien (z. B. XING-Userprofile) generiert, jedoch jeweils unter deren spezifischen Anforderungen wie Qualitätsfaktoren, Aktualisierungsdatum und Seitentyp. Besondere Aufmerksamkeit erhalten außerdem die Privatsphäre-Einstellungen: XING-User können jederzeit selbst bestimmen, ob ihre Inhalte in Suchmaschinen auffindbar sein sollen oder nicht – der Effekt dieser Einstellung muss technisch einwandfrei unterstützt werden.

Ein zusätzlicher Validierungsschritt und ein an das System angedocktes Monitoring sichern die Qualität der Sitemaps und senden Informationen und Fehlerberichte.

Effekte bei xing.com nach Launch der Sitemap-Applikation

Man kann es nicht anders sagen: Der Effekt der neu generierten Sitemaps und Feeds aufs Crawling war enorm und übertraf alle optimistischen Erwartungen. Nach einer deutlichen Stabilisierung der zuvor schwankenden Crawlrate (Anzahl der pro Tag gecrawlten Seiten, Abbildung 7) stieg diese laut Google Search Console und internen Auswertungen im Verlaufe des Folgejahres massiv an. Wenn Crawling die Voraussetzung für Indexierung und Rankings bzw. Traffic ist – hatte das auch darauf positive Auswirkungen?

» Indexierung

Durch den Einsatz der Feeds mit neuem/aktualisiertem Content sowie zu deindexierenden URLs wird frischer UGC nahezu sofort gecrawlt und kann dementsprechend ebenso schnell indexiert werden. Umgekehrt werden gelöschte oder nicht zu indexierende Inhalte sehr schnell deindexiert. Bemerkbar machte sich dies zunächst bei Stichproben, vor allem aber konnte ein starker Rückgang an Userbeschwerden verzeichnet werden – ein klarer und wichtiger Beleg dafür, dass das über Sitemaps gesteuerte Indexierungsmanagement bestens funktioniert. Sicherlich wissen das nicht nur die User zu schätzen, sondern auch Suchmaschinen.

» Rankings/Traffic

Hier war der Effekt tatsächlich am deutlichsten zu spüren: Nach dem Launch der Applikation stieg der SEO-Traffic innerhalb von zwei Wochen signifikant an. Es bestätigte sich hiermit die Vermutung, dass tatsächlich Crawling-Probleme aufgrund der komplexen Struktur vorgelegen hatten: Es gab etliche URLs, die von Suchmaschinen nie erfasst worden waren und demnach auch nie indexiert werden konnten. Auffällig war, dass Bing anscheinend weitaus größere Probleme als Google beim Crawlen der Plattform hatte, denn bei Bing zeigte sich ein weitaus größerer Anstieg von Impressions (Abbildung 8) und SEO-Traffic.

Fazit

Der Einsatz einer ausgefeilten Sitemap-Applikation macht sicherlich nicht für jede Webseite Sinn. Große Seiten mit komplexer, tiefer Struktur, unterschiedlichen oder viel Bewegung in den Inhalten, die auf ein effizientes Handling angewiesen sind, können jedoch durchaus davon profitieren! Nehmen Sie sich daher die Zeit und prüfen Sie die Anforderungen an Ihren Content genau, um ein passgenaues Konzept zu entwickeln. Die technische Umsetzung erfordert gegebenenfalls einiges an Aufwand, lohnt sich aber in doppelter Hinsicht: zum einen für die Unterstützung von Suchmaschinen bei Crawling und Indexierung, zum anderen für die Vereinfachung von Auswertungen, die wichtige Erkenntnisse bringen können. Diesem Thema widmet sich der zweite Teil in der nächsten Ausgabe der Website Boosting. Sie erfahren dann, wie Sitemaps Sie bei der SEO-Analyse unterstützen können, wie Sie Probleme bei der Indexierung identifizieren und wie ein einfaches Monitoring auf Basis von Sitemaps aussehen kann. ¶