On improving the indexing of information systems based on web technologies

S Kratov
Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Novosibirsk, Russia

Abstract. The article describes the approach to the modernization of information systems based on Web-technologies with the aim to improve their indexing by search engines, to increase the interconnection of such systems content, improving GUI in terms of interaction with social services and networks. The most important metrics for statistical data collection systems are described. The set of criteria to evaluate the work of SEO specialists is given.

1. Introduction
The article describes the process of modernization of the information system based on Web technologies in order to comply with modern requirements and standards. As an example of the system being modernized, the informational system for supporting the scientific news site http://academcity.org is used, but all methods and recommendations described in this article can be applied to any other information system accessible to its users via the Web.

The information system for supporting the scientific news site developed earlier was subjected to modernization. With its help, journalists and editors can post on the site and make available to users (readers) the entire set of information materials of the system (news, articles, infographics, videos, surveys, etc.). The complex of basic information system solutions (technical architecture, subsystems) was developed based on content management framework (CMF) Drupal. Drupal is the framework for designing content management systems written in PHP and using relational database as a repository. CMF Drupal is focused on the LAMP platform (Linux OS + Apache + MySQL + PHP language), but other system software options are possible. CMF Drupal has the modular architecture with a compact kernel that provides an API, which can be accessed by modules created by third-party developers. CMF Drupal is free software, protected by the GPL license. Thanks to all of the above, it is very often used to develop Web-based information systems [1-5].

2. Problem definition
After the development of the information system for supporting the scientific news site and the annual period of its use, the need for its modernization became evident, the goals of which were:

- Optimization of the layout of the information system pages and improvement the indexing of the system content by search engines,
- Optimization of the content structure and increasing the interconnection of the information contained in the system,
- Improvement of the GUI in terms of interaction between the system and social services and networks,
- Estimation of the SEO specialists work results on the basis of collected statistical data.
The principles and specific parameters for making such changes in the modernized information system are described below.

3. The optimization of content structure and layout of information system pages

Further, in the text of the article, a "page" means any type of information material in the system: document, news, infographic, video, survey, and the page itself. For optimal indexing of the pages of the system by search servers (Google, Yandex), optimization of their layout and markup is required. It must meet the following criteria:

Uniqueness of the text - the text of all pages of the system should be 70-80% unique.

Keywords in the text - the key words for search engines should be found 5-7 times in different variations in the text of a single page.

Keywords in the title of the page (HTML tags) - keywords should be written in the HTML tags of the title of the page ("TITLE", "KEYWORD", "DESCRIPTION"). Fields should be automatically filled with a set of data common to all system pages as a whole, as well as divided by type of material (news, documents, infographics, etc.) and rubrics. To automate the filling of meta tags in CMF Drupal, you can use the ready-made module Meta tags quick (https://www.drupal.org/project/metatags_quick).

- Recommendations for completing "KEYWORDS". There are approximately 5-6 keywords of the page separated by commas (for search engines). It is desirable that these words are contained in the text of the page.

- Recommendations for completing "DESCRIPTION". There is short description of the page (for search engines). Recommendations for completing this field can be found here - http://help.yandex.ru/webmaster/recommendations/using-meta-desc.xml.

Keywords in the body of the page (HTML tags) - in the body it is necessary to use the tags "H1" and "H2" for the headings, "STRONG" for highlighting the important (key) words of the material. Accordingly do not apply these tags for registration of service forms, "hidden" headings, etc.

Interconnection of system pages - system pages should contain links to each other in the text, so that each page is associated with as many other pages as possible. Also for this purpose you can use:

- Block "Related materials" – links to other pages of the system of the same subject, based on the tags provided by the editor.

- The system archive, displayed on all pages of the system in the form of a calendar. When you select the day of the calendar, the corresponding page with a list of all the materials of the selected day is displayed.

Description of illustrations - for the system materials illustrations, meaningful filling of the fields "ALT" (it is displayed when the illustration is impossible to display) and "TITLE" (it is displayed when the mouse cursor is moved to the illustration, full-size viewing of the document) is necessary.

Recommendations for completing "ALT" and "TITLE". The information entered in these fields will be identified for the search engines with the illustration, used for searching by images. It is necessary to enter in these fields the significant information connected with the material or directly with the illustration. At the same time, the "ALT" is likely to not be seen by the visitors of the system, but the "TITLE" will be visible to them. That is, in the "ALT" it is possible to enter a set of keywords, and in the "TITLE" more meaningful information for visitors. For example, the name of the material, the announcement of the material, the name of the illustration itself, tags. Although it is possible to fill both fields equally. It is recommended that the graphic files themselves be given meaningful names, for example, "Surname.jpg", "Chair.gif", "Holiday.png", etc.

Sitemap - it is necessary to create the system sitemap (tree of all the system's pages) in the form of the standard file "SITEMAP.XML". It should be automatically updated in accordance with the
changes (addition, editing, and deletion of materials) in the system. For automated generation of sitemap, its regular updating and uploading to search engines in CMF Drupal, it is possible to use the ready-made XML sitemap module (https://www.drupal.org/project/xmlsitemap).

RSS – it is desirable to create an automatically updated RSS-feed with headings and announcements of all newly appearing pages of the system.

4. Interaction between the system and social services and networks
In modern conditions, it seems feasible and even necessary to organize the most complete interaction of the system with the most popular social networks and services. This will attract new users to the system; organize notification of users about the appearance of new materials, etc. For this purpose, it is recommended to use the following methods:

Place the "Share" buttons on all system materials, enabling visitors to instantly share a link to the material in their accounts in social networks. It is possible to use ready-made sets of buttons provided by Yandex (http://api.yandex.ru/share/), Share42 (http://share42.com/ru) or something similar.

Organization of signing/logging through social networks/services. In the case of using CMF Drupal, there are modules that implement this functionality, for example, uLogin (https://www.drupal.org/project/ulogin). If case of other development tools, it is also possible to use a number of libraries based on OpenID, etc.

Creation of "representations" of the system in social networks - accounts / groups, in which system materials updates will be announced, as well as other materials not suitable for the main filling of the system format.

5. Collection of statistical data
To evaluation the effectiveness of changes to the system, it is need to install on the site counters from Yandex.Metrics (https://metrika.yandex.ru) and / or Google Analytics (http://www.google.com/analytics). With their help, it is possible to quickly assess the impact of those or other changes introduced, as well as obtain a large set of statistics about the system for any period. Examples of individual metrics that can be useful to administrators and system editors are listed below. As a source of these data, the site http://academcity.org, mentioned in the introduction, was used.

The 200 unique visitors (maximum of 354) visited the site per day on average, which viewed 300 pages (maximum – 534). In total for the specified period the site was visited by 39575 unique visitors, they came 55275 times and looked 83009 pages (fig. 1).

In figure 2 are shown the sources of visitors – in general for the entire period (from 55275 times) there were direct hits (set addresses in the browser, bookmarks, etc.) – 12362, search engines – 35471 (Google – 18000, Yandex – 17500), internal jumps (links to other pages "inside" the site) – 1635, social networks – 3202 (Vkontakte – 2190, Facebook – 688), links from other sites – 2483.

Sex and age structure: 2/3 – men, 1/3 – women; 17% – 18-24 years, 43% – 25-34 years, 9% – 35-44 years, 30% – over 45 years old.

The bounce rate, page depth and time on the site are displayed in figure 3. Page views: 80% of users viewed 1 page, 9% – viewed 2 pages, 2% – viewed 3 pages, 2% – viewed 4 pages. Time on the site: 15% users spent 0-10 seconds on the site, 50% spent 11-30 seconds on the site, 6% spent 31-60 seconds on the site, 10% spent 1-3 minutes on the site, 10% spent 3-10 minutes on the site, 8% spent 10-30 minutes on the site.

Attendance by time of day – evenly from 9:00 to 23:00, the rest of the time is significantly less.

The total number of visits (from 83,000 views in total): 1 visit – 23500, 2-3 visits – 10500, 4-7 visits – 8000, 8-15 visits – 7250, 16-31 visits – 8750, 32-63 visits – 8750, 64-127 visits – 7500, 128-255 visits – 3000, 256-511 visits – 750.
Figure 1. The number of unique site visitors per month.

Traffic source

Figure 2. The sources of the site visitors.
Figure 3. The depth of pageviews and the time spent by users on the site.

Files downloads: 1700 – full-screen viewing of illustrations in documents, 750 – downloading files attached to documents, 550 – downloading files attached to news, 500 – full-screen viewing of infographics.

Maps of links / clicks – most actively users of the site main page click on: the names of the headings (as in the top menu, as below the text of the page), the news list (in the upper left corner), the latest materials in separate columns (those with illustrations) and slider (rewind buttons).

Views from mobile devices – 25% of the total number.

Figure 4. The types of user devices.

The average load time for the page is 1.0 second.
During this period, the site was available for 96.5% of the time.

6. Estimation of the SEO specialists work results
After the implementation of all the above-described recommendations, a further step in promoting the system will be the involvement of specialists in search promotion. Estimation of such specialists work success can be done by:
Increasing the attendance of the system – attendance should be tracked for a sufficiently long period (for example, a month).
Raising the system position in the search engines lists for the interesting queries – for example, earlier on the query "..." the system was shown at 100th position and on 5th now, etc.
Increasing the Yandex thematic index of citation (TCI) for the system as a whole – the frequency of its change is not high, it can reach up to 2-3 months.
Increasing the PageRank of the system's pages from Google – the frequency of its change is also not high.
The speed of system's new pages appearance in the listing of search engines.

7. Conclusion
Directions for further system modernization. At the first stage of the development, the system was created that accumulates and structures all the information received and created by journalists and editors. The system was implemented based on modern technologies, which allows providing its visitors with the maximum possible set of constantly updated information, and organizing a full-value writing and publishing of materials by the editorial group. On subsequent stages of modernization it seems expedient to continue system development in the following directions:
Optimization of system pages for their correct display on the widest possible range of mobile devices. Now, the number of sites users, accessing it through mobile devices is at least 25%. Accordingly, it seems appropriate to develop site versions for various mobile devices (separation depending on the screen resolution and processing power) and their operating systems (iOS, Android, Windows Phone).
Organization of the system materials broadcasting to external sites and services in automatic mode and using manual filtering and sorting of materials, taking into account the peculiarities of a social networks, in order to reach the widest possible audience of potential visitors. Organization of information interaction with other similar subjects sites:

- Organization of electronic mailing lists for formed and constantly updated groups of users.
- Organization of the exchange of links, banners with other similar themes sites.
- Organization of broadcasting outside, as well as the publication on all "related" sites of topical RSS feeds, made up of the materials of all participating sites.

Acknowledgments

The research was supported by the ICMMG SB RAS budget project N 0315-2016-0006.

References

[1] Caron B, Toole D, Wicks P and Miller S 2011 Digital Ocean: building a platform for scientific collaboration and social and media sharing on the Drupal content management system Earth Science Informatics iss. 4 Special Issue: SI 4 pp 191–6
[2] Drouillon P J and Dutre P 2012 How to (re-) build a frequent asked questions site with drupal, an open-source CMS 5th Int. Conf. of Education, Research and Innovation (ICERI 2012) pp 908–12
[3] Sokolova O D and Kratov S V 2016 Information systems for popularization of scientific and knowledge-based software 2016 13th Int. Scientific-Technical Conf. on Actual Problems of Electronic Instrument Engineering (APEIE) – 39281 Proc. 1 (Novosibirsk) pp 519–22
[4] Tramullas J 2013 Content management with Drupal: a review of modules specific to libraries, archives and museums Profesional de la Informacion iss. 5 22 pp 425–31
[5] Zybarev Y and Kratov S 2012 The information support system of corporate software resources development 2012 7th Int. Forum on Strategic Technology Proc. (IFOST 2012) (Tomsk) pp 683–85