Project presentation

The e-Beliana Project

Tatiana Šrámková
The Society for Open Information Technologies
Piešťanská 2991/2, 010 08 Žilina, Slovakia
tana.sramkova@soit.sk

Miloš Šrámek
The Society for Open Information Technologies
Piešťanská 2991/2, 010 08 Žilina, Slovakia
milos.sramek@soit.sk

Viera Tomová
The Encyklopedic Institute of the Slovak Academy of Sciences
Bradáčova 7, 851 02 Bratislava, Slovakia

Summary

This paper introduces e-Beliana, a project of the Encyclopedic Institute of the Slovak Academy of Sciences and the Society for Open Information Technologies, a non-profit organization aimed at the promotion of free and open-source software, open data and open government in general. The main purpose of the collaboration is to publish any current and future content of the Encyclopædia Beliana, the first general Slovak-language encyclopedia, on the internet under a Creative Commons license. In order to accomplish this goal the whole process of preparing the Beliana articles had to be changed substantially. The new web-based editorial system e-Beliana is the topic of this paper.

Key words: encyclopedia, editorial system, open source software, data analysis and conversion

Introduction

The Encyclopædia Beliana is a Slovak-language general encyclopedia. It was first published in 1999 by the Encyclopedic Institute of the Slovak Academy of Sciences and thus far contains 55,000 articles and 12,000 illustrations in eight printed volumes covering the letters A – K. The plan is to publish the last volume in 2031. The abovementioned facts and the rise in popularity of internet encyclopedias inspired the institute to change the current editorial process and make the content of Beliana available on the internet.

Therefore, the institute started a common project called e-Beliana with the Society for Open Information Technologies, a Slovak non-profit organisation aimed at the promotion of free and open-source software [SOIT]. Both parties signed a contract according to which SOIT will create an open source based editorial system for free. This system will make the editorial process more efficient and will allow easy transfer of content to the internet. According to the contract, the Encyclopedic Institute will publish any existing and future content under the Creative Commons SA BY licence [CC].

The old and the new workflow

The existing editorial workflow was designed about 20 years ago using the technical means of that era. It was designed for the production of the printed encyclopedia with no intention to publish its content online. The editorial workflow consisted mainly of exchanging of MS Word documents between authors and editors and among editors themselves. The proofreading was done entirely on paper with the document files being subsequently modified (error prone). Articles were proofread in batches of about 200, with each batch containing articles from different categories. A batch could be proofread by only one editor at a time. This approach resulted in a large time gap (even many months) between the preparation of the original text and the proofreading.

The aim of the new web-based editorial system e-Beliana is to overcome the abovementioned drawbacks and to prepare articles simultaneously for the printed and the internet version of Beliana. The new system should enable independent editing of articles which should significantly shorten processing time. The system should be flexible enough to implement the currently used workflow and to easily allow modifications according to future experience and demands.

The basic requirement of such an editorial system is that in a given moment only one author or editor can modify an article. This requires both horizontal and vertical specification of access rights. Horizontally, access rights must be granted to different authors and editors according to the category of the article. Examples of categories are “Mathematics”, “Zoology” or “German literature” (there are currently more than 600 such categories in Beliana). One editor is responsible for several categories, and there may be several authors for a single category. Vertically, an article flows through a sequence of stages: with author, editor, consultant, etc. Therefore, a user (authors, editors, senior editors, etc.) should not only be able to edit articles in a specific stage but also to change this stage to the next one in the workflow.

Editorial system e-Beliana

Based on the abovementioned requirements we analysed various available open-source solutions. From among them we selected the *Drupal* CMS [Drupal] with its modules *Workbench Access* and *Workbench Moderation* which made the implementation of both horizontal and vertical access rights possible. Subsequently, we implemented a first version of the software, imported the published articles and made the system available to users. Even this preliminary version provided them with incomparably faster search functionality in the texts of published articles as compared to the earlier option – searching in hundreds of MS Word or pdf files or browsing through the printed book. The remaining functionalities of a full editorial system have been implemented gradually since then in close collaboration with the editors. This approach would not have been possible with a proprietary system – in order to try something out, we simply downloaded and installed the software (mainly Drupal modules extending its basic functionality) from the internet, no negotiations and license purchases were necessary.

Several special parts of the software that provided functionality not readily available in Drupal, were later implemented.

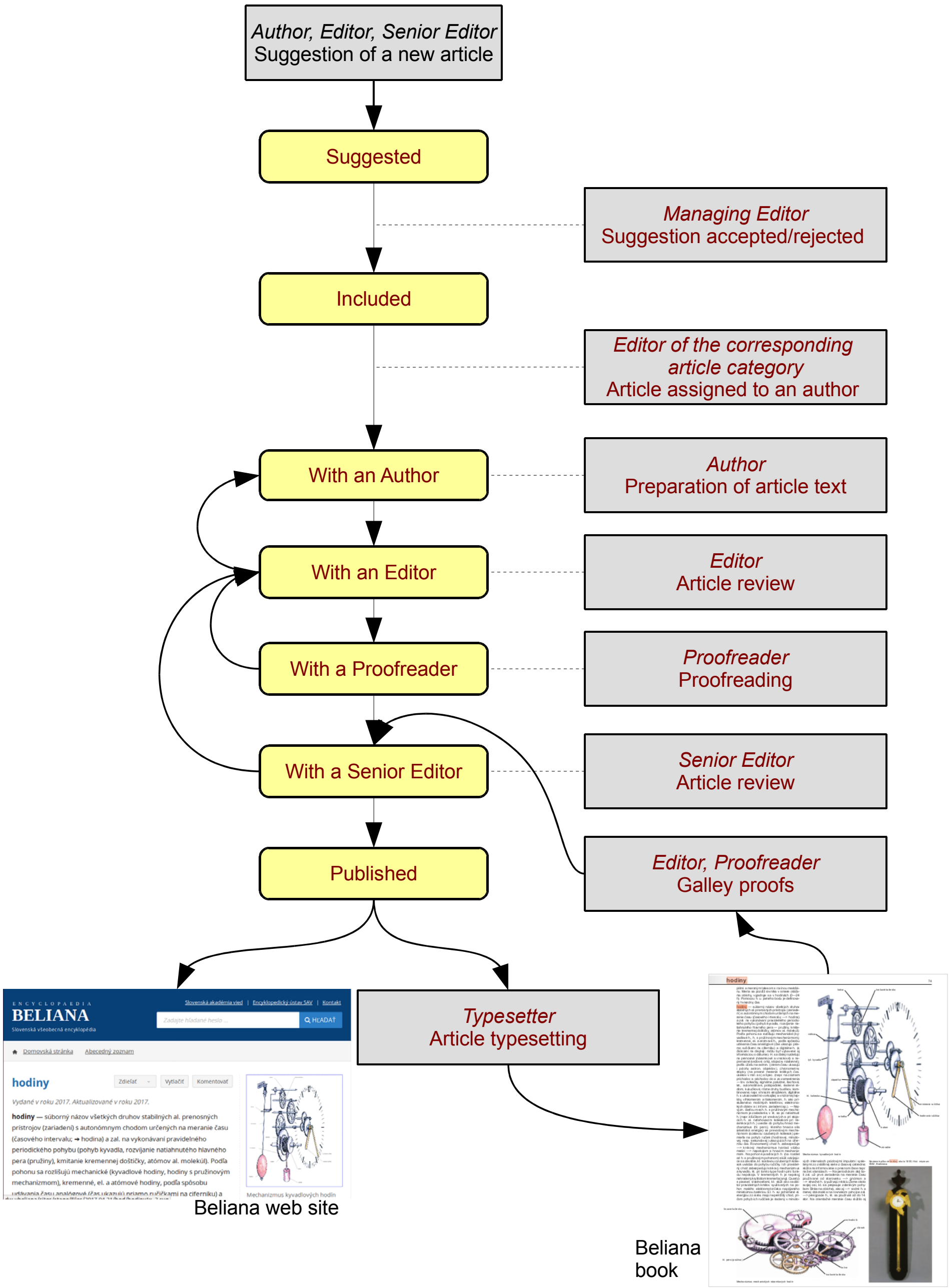
Currently the e-Beliana workflow (Figure 1) uses 27 article stages (e.g. proposed, accepted, with author, with editor, etc.) and 14 user roles (e.g. author, editor, consultant, etc.). Users can search and process available articles in the editorial system by using *views* (implemented by means of the Drupal's *Views* module), which were designed according to the demands of their role. Currently, there are about 20 views which take the access rights provided by the Workbench Moderation and Workbench Access modules into account and about 20 general views which either provide an overview of all Beliana articles to any user or provide users with any necessary special rights For example, editors can access a view which enables them to list articles waiting to be edited or a view that allows them to assign articles to authors for editing.

Articles can be opened for modification from a view. Text can be edited in an embedded editor [CKEditor], which in addition to basic formatting enables users to edit mathematical and chemical formulas in LaTeX notation and to track editorial changes by storing information about the date and author of the change. Formulas are edited and displayed in Beliana using the MathJax tool [MathJax]. The change tracking capability is provided by the CKEditor's LITE module [LITE]. Changes can be viewed not only in the editor window, but in all stored revisions of an article.

Exporting articles to web and printing

Articles which have passed the editorial workflow are ready for publication – in our case both publication on the internet and in book form. A Beliana website (not accessible to public yet) has been developed independently of the editorial system. Contrary to the complexity of the editorial system which uses tens of additional Drupal modules, during development of the Beliana site, we mainly focused on simplicity, robustness and speed. Editiorial system content is synchronized to the public website using the REST API.

While the article text is exported to the internet site without changes, it must be modified for the printed version. The major text modifications are the removal of hyperlinks and the abbreviation of common words and article title occurrences in text. The abbreviation tool is based on the stemming and subsequent



hodiny

Vydané v roku 2017. Aktualizované v roku 2017.

hodiny — súborný názov všetkých druhov stabilných al. prenosných prístrojov (zariadení) s autonómnym chodom určených na meranie času (časového intervalu; ➜ hodina) a zal. na vykonávaní pravidelného periodického pohybu (pohyb kyvadla, rozvíjanie natiahnutého hlavného pera (pružiny), kmitanie kremennej doštičky, atómov al. molekúl). Podľa pohonu sa rozlišujú mechanické (kyvadlové hodiny, hodiny s pružinovým mechanizmom), kremenné, el. a atómové hodiny, podľa spôsobu udávania času analógové (čas ukazujú priamo ručičkami na ciferníku) a

Figure 1. Simplified editorial workflow. Rounded boxes represent editorial stages, sharp boxes represent actions (with user role in italics).

automated inflection of words [Garabík, Radovan] and enables the abbreviation of single words as well as phrases. Beliana books are typeset in InDesign using the MathMagic tool for rendering of LaTeX-based mathematical and chemical equations.

Importing content from existing sources

Each article was imported either from the corresponding text document (published articles) or from a general alphabetic index (future articles). The general alphabetic index is a list of articles to be included in Beliana. It is represented by a set of spreadsheet files containing information such as title and category. The published articles had been created in different versions of the MS Word text editor. Neither the index nor the article document files were edited and maintained with the intention of further machine processing. Both contained numerous random comments and instructions embedded directly in the text. It often became unclear from the point of view of programmatic analysis what was the real text and what was the comment or instruction. The MS Word document files for the first two encyclopedia volumes were not even available; these texts had to be extracted from PageMaker (volume 1) and QuarkExpress files (volume 2).

The development of import tools was not straightforward and required many cycles in which they were refined. This was caused not only by the aforementioned random comments and instructions, but also by the fact that we strived to extract certain information from loosely structured human readable text, as was the case of detection of names and surnames or dates of birth and death.

The import tools were implemented using *bash* as a basic scripting tool, *LibreOffice* for conversion of files in MS Office formats for further processing, the *aspell* spellchecker for detection of misspelled words in the Slovak and Czech languages and the stream editor *sed* for batch replacement of misspelled characters and words. Our own tools were written in *python*.

Among others, the import tools included the following operations:

1. *Correction of incorrectly recognised characters*. The oldest documents were from the 90s and included special and accented characters which were not correctly recognised by today's software.
2. *Text segmentation in articles and detection of the article title*.
3. *Detection of personal names and surnames*. In articles related to persons, surnames were used as an article title (i.e., it was typeset in bold). In the internet version of Beliana we prefer to have article titles with both the name and the surname. In order to detect the first names we implemented a rule based system, taking into consideration different kinds of information (for example, the title was followed by a comma, the article was not in a category related to geography etc.)
4. *Detection of the article category*. Information about the category of the article was available from two sources: the general article index spread-

sheets and one of the early versions of the article texts. A rule-based approach taking advantage of fuzzy matching was used, in order to detect the category of two or more articles with the same name or articles with multiple categories created by merging other articles.

5. *Detection of references to other articles*. A reference to another article was marked in the Beliana text by an arrow which was followed by the title of the referenced article. This title, however, was often inflected and abbreviated and was often not separated from the subsequent text. Therefore, we always considered all words following the arrow till the first non-letter and non-space character (except for a dot) to be title candidates. Words of this string were subsequently lemmatized [Garabík, Radoslav] and an article with the most similar title was looked up in the list of all articles.
6. *Detection of illustration captions*. Illustration captions were included directly in the article text and were separated from it by using the words "Illustration text". Since these were often abbreviated and formatted in different ways, not all captions were detected.

Using the analysis tools, we processed and prepared about 270 document files with the text of the published articles (more than 6 mil. words, 56,000 articles and 12,000 illustrations) and 12 spreadsheet files of the general article index with more than 110.000 entries for import. The analysis tools developed in this process were not perfect, but when compared to purely manual processing they saved a substantial amount of work.

Conclusion

In this paper, we have briefly introduced the e-Beliana editorial system based on the Drupal CMS and its modules. The open-source character of Drupal was the most important feature for a successful realisation of the project. The source code is available in the GitHub repository of the Encyclopedic Institute of the Slovak Academy of Sciences [GitHub]. As the new editorial system has only been used for 3 months, it is too early to summarise its advantages and drawbacks. The public web site will be available in the beginning of 2018.

References

Garabík, Radovan. Slovak morphology analyzer based on Levenshtein edit operations. In: Proceedings of the WIKT'06 conference. Laclavík, M.; Budinská, I.; Hluchý, L. (ed.) Bratislava: Institute of Informatics SAS, 2006, 2 – 5

SOIT. About us. http://soit.sk/sk/in_english/who-we-are (June 30, 2017)

CC. Creative Commons. https://creativecommons.org (June 30, 2017)

Drupal. http://www.dupal.org (June 30, 2017)

CKEditor. http://ckeditor.com/ (June 30, 2017)

MathJax. https://www.mathjax.org/ (June 30, 2017)

LITE. https://www.loopindex.com/ (June 30, 2017)

GitHub. https://github.com/enu-sav (June 30, 2017)