OPENOFFICE.ORG: EVERYDAY FREEDOM NOW IN SERBIAN AS WELL

Goran Rakić

Faculty of Mathematics, Belgrade

Abstract: In this paper we mark the disadvantages of the use of proprietary office software formats and present an alternative internationally standardized Open Document Format (ODF). The most popular free office suite OpenOffice.org and its version 3.0 is briefly presented in the second section. The third section of this paper deals with the Serbian localization of this office suite, which was created through the cooperation of Faculty of Mathematics in Belgrade and the Serbian OpenOffice.org community. In this section the localization work flow is depicted and the most important issues, as perceived by the author working on Serbian free software localizations for many years, are stressed.

1. Open document formats

Due to the unfair business model of software vendors, we made choice of one specific type of office suite software that effectively locks all users' future documents for only one vendor and only one solution. Although data is still owned by the user, documents written using a non- standardized closed proprietary format are unreadable without software used to produce them. The proprietary document format is a secret known only to the software vendor, which they will use later to leverage their market advantage.

Users are not able to switch to other software solutions since with them they will not be able to open their documents. Even upgrading to a new version or migrating to a new release can make access to old documents impossible.

In an age in which digital data is more and more rapidly replacing paper documents, people are becoming aware of this proprietary problem. Many even have painful personal experiences with it.

Despite the fact that technological developments and previous practices have enabled the lifelong storing of digital data, which is necessary to handle the exponentially growing amount of information ($4 \cdot 10^{19}$bit only in 2003, and doubling every three years, according to (Lyman & Varian 2003)), the average lifespan of a proprietary document format is just a few years.

The most used definition of "open format" is that it is a standard that meets the following requirements (Wikipedia – Open Format):

1. A publicly available specification released under a license that puts no restrictions on the right to study, modify and implement and that also requires the publication of all modified versions of specifications under the same conditions. It is not allowed to have discriminative clauses in respect to the implementation of any published specification either by patent regulations or any other.
2. An open standard management in the form of an independent working group with equal rights for the membership of everyone and a working procedure that favors no interested party.
3. The available free implementation of published specifications, usually as an open code software product released under a free software license or as under the public domain.
4. The requirement to have at least two independent implementations conforming to a standard.

The only standard for office document formats, including text documents, spreadsheets and presentations as well as components for rep-

resenting included sub documents that meets all of these four requirements and which is ready for use today is the OpenDocument Format (ODF). This format is standardized by ISO as standard number 26300:2006.

Standard management is in the hands of the nonprofit organization OASIS (*The **O**rganization for the **A**dvancement of **S**tructured **I**nformation **S**tandards*). The technical committee responsible for the development of the standard consists of leading software vendors (Adobe, Google, IBM, Intel, Microsoft, Novell, Sun Microsystems and others), as well as engaged individuals and representatives of academia.

Technically speaking, the OpenDocument Format is actualized as a language based on XML, packed without any loss of information using Huffman coding compression (ZIP DEFLATE format). Its XML representation uses other internationally recognized standard open formats for representing subdocuments widely used in other software solutions: MathML[1] for the representation of mathematical expressions, the *Dublin Core*[2] subset for representing metadata, *SMIL*[3] for representing animations in presentations, etc. This way, implementation relies on existing solutions while also keeps the size of format specification at a tenth of the size (or less) of competing proprietary formats.

Due to the fact that the data format is based on XML, and as such is a plain structured text, information readability is preserved even in case of data corruption on storage medium or in case of transmission errors.

Using of open document formats like the OpenDocument Format is a necessary step toward software vendor interoperability and a key factor to obtaining lifelong digital records.

In the context of e-Government, through requesting the usage of high interoperability standards and open formats, no specific vendor is favored. This enables anyone communicating with the government electronically to make their own choice of software they will use. Data records remain transparent and available to everyone. It is clear that one cannot speak about a democratic society in context of e-Government if these requirements are not fulfilled.

The OpenDocument Format was established as a norm by legislative acts for the e-Governments of Norway, Belgium, Finland and France, while other countries adopted a recommendation for its usage (Germany, Japan, and Slovakia) (ODF Alliance). In July of 2008, NATO adopted this format as a mandatory standard for its members (NISP2).

The adoption of open formats with their free and available implementations allows for the undisturbed knowledge transfer between developed to underdeveloped countries in an information society. It is therefore understandable why the wide adoption of open formats contributes to a solution of the problems presented in the UN Millennium declaration (UN-55/2) and acknowledged by the agenda of the World Summit on Information Society WSIS held in Tunis in 2005 (WSIS05).

With support from industry leaders and also with high quality technical solutions, a widening use of the OpenDocument Format has the potential to change today's unfair business practice and offer users a safer working environment that is less liable to risks.

2. OpenOffice.org 3 for the Users

Simply, OpenOffice.org is the most popular free office software suite.

The suite consists of a text processing application, spreadsheet, an application for preparing and showing presentations, an application for writing mathematical formulae, an application for vector drawings and an application for database creating and management. All of these provide complete software support for everyday office work.

Its software can be installed on all popular computer types and it supports operating systems Microsoft Windows, GNU/Linux, Apple Mac OS X and Solaris. Installation can be downloaded from the project web site: http://sr.openoffice.org.

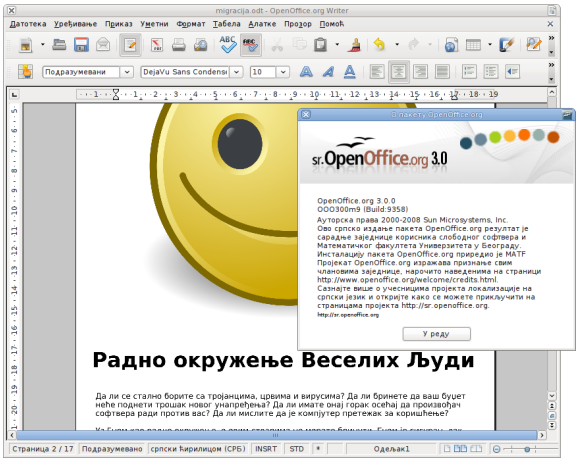

Figure 1: OpenOffice.org 3 Writer — text processing application localized into Serbian

OpenOffice.org is developed as free software, which means that it can be used, copied and distributed unconditionally without any licensing costs. The complete source code is available with permission to customize it and republish its customization under the terms of the GNU Lesser General Public License (GNU LGPL), version 3 as published by the Free Software Foundation[4]. ODF is its default document format but other popular office document formats are supported as well.

It should be noted that although ODF is the default format of OpenOffice.org, it is an open standard that is used by many other applications, both proprietary (closed) and open and free.

The OpenOffice.org project, which turned eight several months ago, gathers programmers, industrial partners developing their own solutions based on OpenOffice.org and worldwide users from the sectors of business, government or education, as well as individual home users. They all support development in order to improve the suite and attract more users. One can find more information on the project's web site of how to contribute and join the project.

There are many new features in the release of 3.0, such as support for PDF importing that allows minor document changes or a new linear solver for operational research in its spreadsheet application. There is also a new color scheme, a refreshed look and support for exporting documents to the Internet, a Wiki or Weblog.

Released under the motto: "All three: Great Software, Easy to Use and Free", OpenOffice.org is empowering everyday productivity for millions o users worldwide.

3. The Localization Process

Graphical user interface for OpenOffice.org has been localized into more than 75 world languages.

The Faculty of Mathematics at University of Belgrade and the Serbian Free Software Community created a localization of OpenOffice.org into the Serbian language, using both the Cyrillic and Latin alphabets. The work was done in the scope of the free software localization project which the Ministry for Telecommunication and the Information Society of Republic of Serbia started in December 2007, the localization is based on two previous localizations from 2005 and 2002 which had been abandoned in the meantime.

The localization material consists of more than 28,000 translation units, that is, unique messages from the application interface, or more than 96,000 words. The work on this project resulted in the publication of complete localizations of releases 2.4, 2.4.1 and the latest release 3.0 (as of November 2008). In January 2009, the complete localization of future release 3.1 was prepared and all errors reported regularly are being corrected.

The Style Guide, with its recommendations for translators, has been published on the project collaboration web site (http://ooo.matf.bg.ac.yu) and its publication has been accepted by members of other localization projects for free software products that are being developed by Community.

The Guide states (Guidelines): "The goal of software localization is to transfer both the direct meaning, and associations that can help the user

to link a concept being translated to those that do not belong to computer world. It is intended for users that would like to use a computer in the Serbian language."

A main benefit of the localized environment's guide points is "For users that do not know English on a proficient level, a localized user interface removes all the barriers impeding the full understanding of software tools, thus making, learning easier and directly improving productivity."

All discussion is done in public on the project mailing list dev@sr.openoffice.org, which is open for everybody willing to improve in localization or to participate in the process.

The complete review and proofreading of Serbian localization was done for the first time during the 3.0 release cycle. More than 4000 mistakes have been spotted, founded and corrected during this process, and better quality has been created. The most common errors include incorrect word order, too frequent use of verbal nouns which is not common in Serbian language, errors in adjective-noun congruence and inconsistencies in terminology.

The biggest problem in software localization in Serbian is the fact that stable and widely accepted computer science terminology does not exist. In order to solve this problem or at least to alleviate it, translators of free software have initiated an open glossary project available at http://recnik.prevod.org. Recently a new computing dictionary project was launched called „Сверачунарски појмовник"(The Almighty Computing Dictionary) which is available at http://nedohodnik.dyndns.org/sverapoj. There is also an open mailing list (sorta@googlegroups.com) where discussion about this dictionary takes place.

It is not uncommon for translators of proprietary software as well as many users to decide to use descriptive translations instead of proper terminology, and to set aside localization with disdain for novice users. Descriptive translations exclude any possibility to use the Serbian language for advanced computing, and a wider adoption of translations is bound to fail.

As soon as a "driver" becomes "management software" (in Serbian „upravljački program"), the problem is sure to appear with the translation of messages such as "Driver management program" or "Managing drivers". Translators then have to use different long paraphrases, which themselves makes the wider adaptation of a descriptive translation impossible.

The only acceptable solution is to map terms into terms. Terms can be imported from general language by their specialization into computing context (ex. *chat-ćaskanje, host-domaćin, proxy-posrednik*), adopted foreign words (usually English words, but not always necessarily, ex. *font-font, to click-kliknuti*) or new words coined following the original etymology (ex. *podcast-podemitovanje*, *thumbnail-sličica*).

When translating messages, special attention has to be paid to agreement with numerals since there are in Serbian several different plural forms. This problem is solved by using different plural forms depending on the number value. Four plural forms in the Serbian language can be selected with following expression in the C programming language:

```
plural=n==1? 3
    : n%10==1 && n%100!=11 ? 0
        : n%10>=2 && n%10<=4 &&
            (n%100<10 || n%100>=20) ? 1
                : 2;\n
```

Message "*%u folder"* (Plural "*%u folders"*) will be translated with four messages: „*%u fascikla*", „*%u fascikle*", „*%u fascikli*" and „*%u fascikla*". According to an expression they will be used for the values:

0) 21, 31,...
1) 2, 3, 4,...
2) 5, 23,...
3) one

The choice of the right gender form for some verb forms is not yet solved, therein translators try to use verb tenses that do not inflect in gender

like present and aorist (ex. "*%s wrote"* where %s is the author's name can be translated with present form "*%s piše"* or aorist form "*%s napiše"* instead of perfect form "*%s je napisao/napisala"* which has to agree with the gender of *%s, which* is in general unknown).

Similar problems exist with noun inflection, specifically when an object has to be automatically inserted into a message that should be in a particular case (ex. "More about %s", Serb. „*Više o %s*"). Another problem is also the translation of Saxon Genitive (ex. „%s's Note") into a Serbian possessive adjective.

The Serbian KDE translators group[5] has offered their best solutions to the second problem by adding scripting to translation loading so that a specific translation variant can be loaded into the runtime[6].

Today it is a general practice to do localization in the Cyrillic alphabet and to automatically transliterate it into the Latin alphabet. There are some style and spelling differences between these two alphabets that should be taken care of (ex. X, Q, W and Y do not exist in the Cyrillic alphabet, but can be used in acronyms in Latin script localization, a hyphen should not be used for inflection endings if a noun is written using the original Latin orthography, etc.). Most Serbian localization projects today neglect these subtleties.

Tools for the automatic translation of quality control such as terminology consistency, spellchecking or the detection of punctuation and style errors have recently been used more in Serbian free software localization projects and, for this reason, better quality is achieved with minimal effort. Some of the tools used are created specifically for the Serbian language, such as pology[7] software created by Časlav Ilić. This software provides many useful tools and a Python framework for developing new scripts for processing files with translations. Some of its features include terminology, consistency and spell checking, customized transliteration, accents and capitalization style, checking and limited support for mapping Ekavian to Ijekavian pronunciation, grammar and style checking, etc.

Apart from terminology, another large problem in localization is the fact that context information is frequently missing during translation. This can lead to a mismatch in gender between the adjective part and the noun part of a message if these parts were input in material separately (ex. The adjective "Automatic" is used together with the noun "Style", but these two messages are separated in the translation material).

The only solution to this category of errors is the live testing of localized software. Labeling messages with alphanumeric prefixes for their easier recognition and selection in the translation material proved to be very useful. Following this approach, a version for the testing of software being translated has been prepared which presents this short code with each message. When an error is detected in the regular localized software, this special version can be loaded and by following the same steps in the user interface, one can easily reproduce the same message and, with the help of this message code, locate it in the translation material.

An automatic checker that would point out occurrences of "English style" word order in translations would significantly improve the translation quality, however the development of such a checker cannot be expected in near future.

4. Conclusion

The Office suite OpenOffice.org can address all user needs. Long term further development is ensured as world leading software companies are beginning to base their new products on OpenOffice.org.

Thanks to the support from the Ministry for Telecommunications and the Information Society of the Republic of Serbia and cooperation between the Faculty of Mathematics in Belgrade and the Free Software Community, the Serbian localization of this software has become available.

As OpenOffice.org is free software, there are no restrictions for its customization. As a result, in the scope of the localization project this suite has been enhanced by Serbian spell-check support, support for list enumeration using letters of the Serbian Cyrillic alphabet, Serbian hyphenation and there are also add-ons for transliteration from Serbian Cyrillic into the Serbian Latin alphabet and vice versa and an add-on for spelling numbers with words.

Commitment to open formats and using the OpenDocument Format as a default allows users the opportunity to choose independence and achieve portability of their documents.

At this point it is necessary to say that, during a past few years, free software localization projects have been done independently by the Community, because of a lack of interest both from academia and language professionals. Many individuals involved had, thus, to abandon their main computing interest and gain strong leverage and experience in localization.

That is why guides like these prepared by the OpenOffice.org localization project and tools like those developed in the KDE translation project are very important as they provide knowledge transfer between individuals involved in these projects.

One can note with great pleasure the increasing interest of translators of proprietary software into activities undertaken by the Free Software Community.

The cooperation of all interested parties is necessary in order to establish Serbian computing terminology, and every step towards this goal is valuable.

[1] Math Markup Language, W3C recommendation http://www.w3.org/TR/MathML2/

[2] Dublin Core Metadata Element Set (DSEC), ISO 15836:2003 and IETF RFC5013.

[3] Synchronized Multimedia Integration Language, W3C recommendation http://www.w3.org/TR/2005/PR-SMIL2-20050927/

[4] http://www.gnu.org/licenses/lgpl.txt

[5] http://sr.l10n.kde.org

[6] http://sr.l10n.kde.org/zihzah.php

[7] http://websvn.kde.org/*checkout*/trunk/l10n-support/pology/doc/html/index.html

Bibliography

Lyman, Peter and Hal R. Varian. 2003. How much information? 2003. UC Berkeley, School of Information Management and Systems. http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/

Wikipedia – Open Format. Wikipedia, Open format article, version dated December 10th 2008. http://en.wikipedia.org/wiki/Open_format

ODF Alliance. ODF Alliance 2008 Annual Report, Annex A, p. 15-26. http://odfalliance.org/resources/Annual-Report-ODF-2008.odt

NISP2. NATO Interoperability Standards And Profiles, Volume 2, Section 3.4 http://nhqc3s.nato.int/architecture/_docs/NISPv2/pdf/NISP-Vol2-v2-internet.pdf

UN-55/2. 55/2. United Nations Millennium declaration, resolution adopted by the UN General Assembly, September 18th 2000.
http://www.un.org/ga/59/hl60_plenarymeeting.html

WSIS05. Tunis Agenda for the information society, November 18th 2005. http://www.itu.int/wsis/docs2/tunis/off/6rev1.pdf

Guidelines. Језичко-стилски приручник за преводиоце пакетаOpenOffice.org,радноиздање31.јул2008.године (Style Guide with linguistics recommendations for OpenOffice.org translators into Serbian language, working copy, July 31st 2008.) http://ooo.matf.bg.ac.yu/prirucnik_ooo.odt