An Open Access Repository for the *Workshop de Software Livre* (WSL)

Filipe de O. Saraiva[1] , **Antonio Terceiro**[2,3]

[1] Escola de Engenharia de São Carlos – Universidade de São Paulo (USP)
Av. Trabalhador São-carlense n. 400 – 13566-590
São Carlos – SP – Brasil

[2]Cooperativa de Tecnologias Livres (COLIVRE)
Salvador – BA – Brasil

[3]Laboratório Avançado de Produção, Pesquisa e Inovação em Software (LAPPIS)
Universidade de Brasília
Brasília – DF – Brasil

filipe.saraiva@usp.br, terceiro@softwarelivre.org

Abstract. *This paper describes the design and implementation of an open access, free of charge repository for WSL. This repository is available at http://wsl.softwarelivre.org/ and contains all the papers from all the editions of WSL since its inception in the year of 2000. The source code of the repository is licensed under the GPLv3, and the papers published under the Creative Commons Attribution-NoDerivs 3.0 Unported (CC BY-ND 3.0) license.*

1. Introduction

WSL is an academic workshop collocated with FISL (Portuguese acronym for International Free Software Forum), the largest Free Software conference in Brazil. Its goal is to showcase free software-related work developed by universities, research centers, companies and hackers. WSL has been held together with FISL since its very first edition in the year of 2000, what at the time of writing means it already had 15 editions.

In part because traditionally WSL hadn't had a fixed set of core organizers, there was never a fixed location for its papers in digital format. The FISL organization stopped producing a paperback version of its proceedings some years ago. This led to a situation where a large amount of the papers from previous WSL editions were not easily accessible, and accessing them required trips to local university libraries.

The last decade has seen the emergence of the Open Access (OA) movement. OA is a social, political and technical movement with the purpose of releasing for public access in the internet, free of charge, the scientific literature, data, software and other resources utilized and produced during the process of creating knowledge. This includes, but is not limited to, scientific research and academic development. OA is usually brought forward by scientists, professors, researchers, publishers, scientific societies, activists, and others.

There were several different OA initiatives in the past, but it is common to recognize the Budapest Open Access Initiative (BOAI), 2002, as the first formal event targeted at spreading the advantages of OA to authors – like more visibility and impact of scientific research – and to the society – e.g. removing the access barriers to research literature will "accelerate research, enrich education, share the learning of the rich with the poor and the poor with the rich", and more. BOAI released a document, like a manifesto, listing the motivations and characteristics of OA and produced the most accepted definition for the term, reproduced below [Chan et al. 2002]:

> By "open access" to this (peer-reviewed research) literature, we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited.

BOAI also defined in that document two strategies to achieve OA for scientific literature: self-archiving of papers and the creation of OA journals.

Ten years after BOAI, a new document named BOAI10 was released [BOAI 2012]. This text is more accurate than the first because it tries to establish a guide of "good practices" that the OA community has learned in the previous 10 years. For example, self-archiving and OA journals are now known as green and gold OA, respectively; there are now recommendations for universities and research centers about institutional OA repositories, recommendation of licenses to be used in the papers, how to make a sustainable OA repository and what are the requirement repository needs to have.

It is now possible to see the impact of the OA community work in the increasing number of new OA journals released in last years, the creation of regulations by some research funding bodies that mandate their scientists to publish papers in OA journals, the creation of optional OA subscription for authors in mainstream scientific publishers, etc [Laakso et al. 2011]. In fact, OA is now a recommended policy by different countries and policy agencies, like UNESCO [Swan 2012].

Free software and OA share a common goal for different type of products: source code for the former, and research papers, data and tools for the later. With WSL being a link between the two communities, it made sense for us to create an Open Access repository for WSL.

This paper describes the design and implementation of such digital repository for WSL, and the work that was performed in order to recover the digital versions of all papers since the beginning of the workshop.

There is already some software available for papers repository management, of which the most well known is Open Journal System (OJS) [Muir et al. 2005]. OJS is a we application and as such requires a certain work to install, configure and maintain. That includes installing the application itself, installing a database server, configuring database access in the application, and finally configure a web server. Our work proposes lightweight OA repository management software where content is exported as static HTML pages: this requires minimal to no configuration at all, since most researchers already have access to web space they can upload static content to). The repository configuration uses a markup language easy to understand for non-programmers.

The code of repository is free software licensed under the GPLv3. Its contents (the full set of papers) is licensed under the Creative Commons Attribution-NoDerivs 3.0 Unported (CC BY-ND 3.0) license, with full, unrestricted access (i.e. no pay wall). The source code is available at https://gitlab.com/wsl/repository/.

This choice of licensing and access model was made by a few reasons:

- to democratize the access to the content published in WSL, which after all, is about Free Software. Not having an Open Access knowledge repository would be a serious philosophical contradiction.
- to create a solution that can possibly be enhanced, generalized, and adopted by other publications looking into building its own independent open access repository.

The remainder of the text is organized as follows. The next section describes the work done to recover all papers from previous WSL editions. Section 3 presents the design of the repository, the metadata structure, and the build process of the resulting website. Section 4 makes presents a tour of the repository features. Finally, Section 5 presents conclusions and points some improvements and future work.

2. Recovering papers from all previous WSL editions

Recovering all WSL proceedings published by the previous editions was a real "internet archeology" job.

In the beginning we had the proceedings in printed book format, but in order to build an internet repository we needed the papers in digital format. There were several previous proceedings scattered in different websites around the web, so we collected those papers to be used in the project.

With some proceedings still missing, we tracked down their committee members and asked them for the digital versions. In parallel, we searched the tools utilized in the paper submission process. Several editions could still be found found in SBC-JEMS (Brazilian Computer Society – *Sociedade Brasileira de Computação* – Journal and Event Management System), utilized in different WSL editions. A wiki page was created to manage this work.

The search for ancient WSL call for papers led the authors to different webpages in free software media websites, old messages in mail lists, and sometimes we needed to use the Internet Archive Wayback Machine to see information in old versions of webpages. From this research we found committee members of all previous editions, and using their names we found their current e-mails and made contact with them.

Some committee members contacted still had the digital version of the proceeding and sent it immediately to the authors. Others didn't have it anymore, and in these cases we needed to scan the proceeding from the paperback versions.

In terms of numbers, from the current 15 WSL editions, we had 9 in digital format. We contacted 6 committee members from the editions missing the digital version of the proceedings. 4 committee members sent to us the digital version. We needed to scan 2 missing proceedings.

With the proceedings in hands, it was time to do some preprocessing work in order to provide the basis for the WSL papers repository.

Some proceedings, even if in digital format, didn't have individual files for each paper, i.e. they were all a single PDF file. Some of these had two physical pages per PDF page. In both cases we used `ghostscript`[1] and `pdftk`[2] to extract the pages and generate the final PDFs.

In the case of the scanned proceedings, the books were available as a large number of JPG files, one for each page of the book. We used a Ruby script that took as input the list of files for each paper (fortunately the JPG files were named sequentially) and used `ImageMagick`[3] to merge all pages of a paper into a PDF file. The script used for each edition was slightly different, and is available in repository source code under the folder for that particular edition, usually called `import.rb`.

After preprocessing the proceedings files, it was time to extract the data from the papers to create the metadata of the proceedings. Metadata was extracted from the papers that were already in proper PDF files by copy-and-paste. In the case of PDFs put together from scans, we used a free software OCR (Optical Character Recognition) tool called `tesseract`[4]. It provided good results and saved some typing.

In both cases some manual work was realized in a few cases where the copy-and-paste process was not successful or when latin characters were not properly recognized.

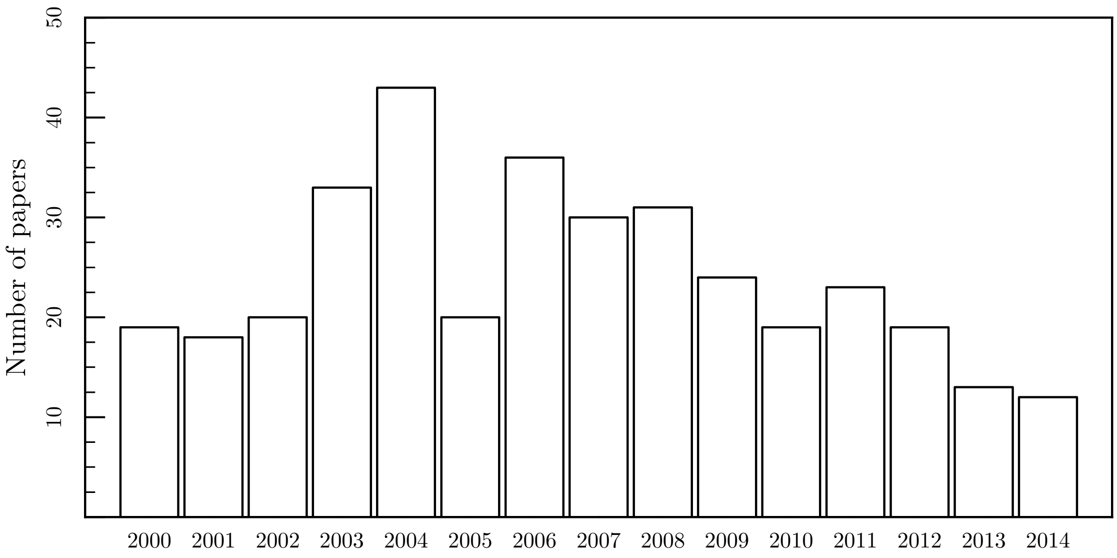


Figure 1. Number of papers by WSL edition

3. Design of the repository

The repository was developed as a website containing different pages for each WSL edition and their respective papers. There is also a search page, which provides an unified vision of the complete data set of papers for all editions.

[1] ghostscript website: http://www.ghostscript.com/
[2] pdftk website: http://www.pdflabs.com/tools/pdftk-the-pdf-toolkit/
[3] ImageMagick website: http://www.imagemagick.org/
[4] tesseract website: https://code.google.com/p/tesseract-ocr/

The website pages are generated from metadata files for each WSL edition. This metadata provide information for each paper, for example, title, authors, abstract, and more. A script will read these metadata and will generate the HTML files of the website, just like other static website generation tools.

The metadata is stored in the YAML format, a human-friendly and easy to understand data description format. An example of the metadata structure utilized in the repository is presented bellow:

```
papers:
  - title: "Paper title"
    abstract: "This paper ..."
    file: "paper1.pdf"
    code: 1
    authors:
      - name: First Author
        institution: XXXX
      - name: Second Author
        institution: XXXX
  - title: "Another paper"
    abstract: "This paper ..."
    file: "paper2.pdf"
    code: 2
    authors:
      - name: Only author
      - institution: YYYY
```

From this metadata and the paper PDF files, the repository will produce a static HTML website. This brings a few advantages:

- the hosting requirements are extremely low. All that is needed is a web server capable of serving static content.
- the repository maintainers do not need to care about performance issues, since the only code that is executed when a user visits the repository is the already highly-optimized web server code.

The directory structure of the repository files is important in order to generate the website correctly. Each metadata and their respective papers need to be in a separate folder related with each WSL edition. The directory structure of the repository looks like the example below.

```
.
+-- Rakefile
+-- templates/
+-- assets/
+-- 2000/
|   +-- data.yaml
|   +-- paper1.pdf
|   +-- paper2.pdf
|   +-- (more papers)
+-- (more years)
```

Each WSL edition will have its own folder with the respective metadata file (named `data.yaml`) and each paper presented in that edition in PDF format.

After the build process, the directory structure of the website will be as below.

```
.
+-- index.html
+-- search/
|   +-- index.html
+-- assets/
+-- 2000/
|   +-- 0001/
|   |   +-- index.html
|   |   +-- download/
|   |   |   +-- index.html
|   |   +-- paper1.pdf
|   +-- 0002/
|   |   +-- index.html
|   |   +-- download/
|   |   |   +-- index.html
|   |   +-- paper2.pdf
|   +-- (more papers)
+-- (more years)
```

The process will generate a folder for each WSL edition and a sub-folder for each paper presented. For example, to *WSL 2000*, the build process will create the folder *2000/* for the edition and the folders *0001/* for the first paper of *WSL 2000*.

For each paper folder in the website, there are three others additional files: the *index.html* will present the data of the paper; *download/index.html* has the code to count the number of downloads that article; and the respective paper in PDF file. In the case of *2000/0001* paper, it is the *paper1.pdf* file.

The website is generated from the metadata files combined with the page template files available. There are 4 different page models in `templates/` folder: `index.html.erb` for the initial page of the repository and for each workshop edition initial pages; `paper.html.erb` for each specific paper information page; `download.html.erb` for download paper pages; and `search.html.erb` for search page.

During the build process, the script in `Rakefile` will look for all metadata in the directories and sub-directories. From it, the script will fill the template pages, add the data obtained from metadata files, and create pages for each WSL edition, paper information, download information, and search.

The repository source code was developed in Ruby for the website generation. Another languages were utilized in others parts, for example, the page templates were developed in HTML, JavaScript is utilized in all dynamic parts of the website (for example, to show paper metrics, paper search, and more), the committee information utilizes markdown, and the metadata, as mentioned before, uses YAML.

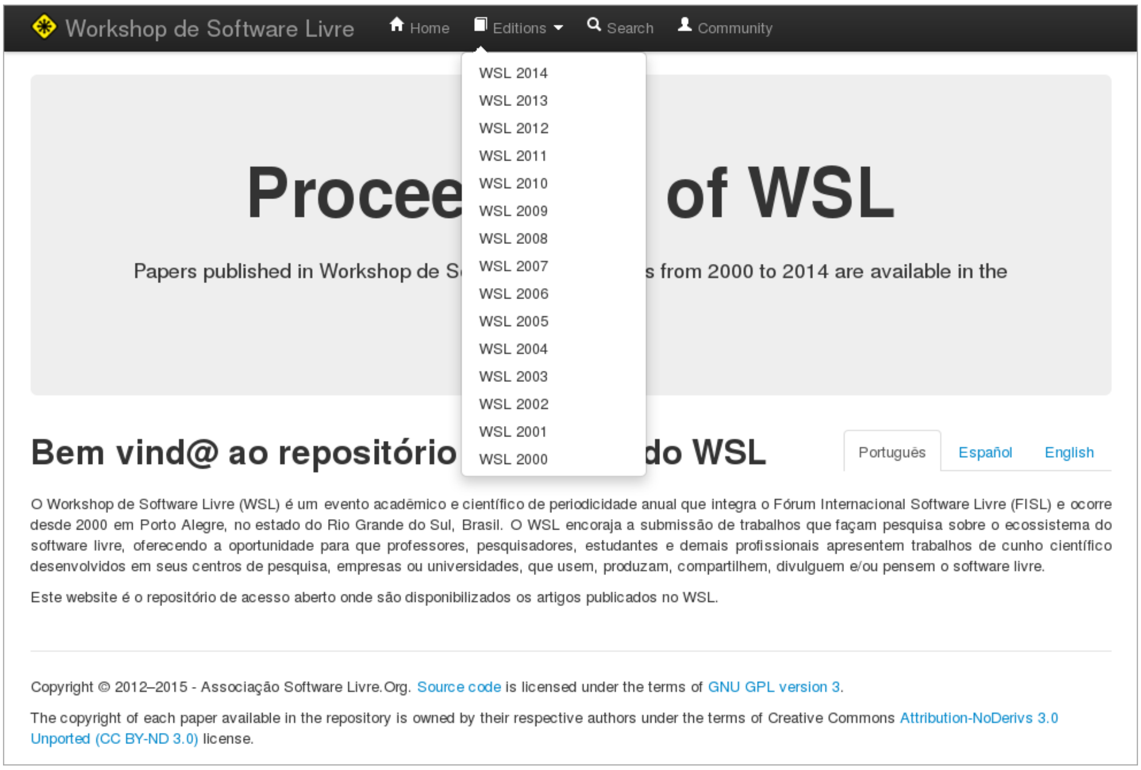
To generate the website repository, one needs to run `rake` in a terminal, from inside the root directory of the repository source. The website will be created in a subdirectory called `public/`. Now, you can run a simple web server pointing to the `public/` directory to browser the repository, or copy the `public/` directory to a web server. It is also possible to open each HTML file of the repository directly in a web browser.

The requirements to build the repository are:

- Ruby
- erubys (Ruby gem)
- yaml (Ruby gem)
- rake (Ruby gem)
- pandoc[5]

4. A tour of the repository features

4.1. Home page and navigation bar



O Workshop de Software Livre (WSL) é um evento acadêmico e científico de periodicidade anual que integra o Fórum Internacional Software Livre (FISL) e ocorre desde 2000 em Porto Alegre, no estado do Rio Grande do Sul, Brasil. O WSL encoraja a submissão de trabalhos que façam pesquisa sobre o ecossistema do software livre, oferecendo a oportunidade para que professores, pesquisadores, estudantes e demais profissionais apresentem trabalhos de cunho científico desenvolvidos em seus centros de pesquisa, empresas ou universidades, que usem, produzam, compartilhem, divulguem e/ou pensem o software livre.

Este website é o repositório de acesso aberto onde são disponibilizados os artigos publicados no WSL.

Copyright © 2012–2015 - Associação Software Livre.Org. Source code is licensed under the terms of GNU GPL version 3.

The copyright of each paper available in the repository is owned by their respective authors under the terms of Creative Commons Attribution-NoDerivs 3.0 Unported (CC BY-ND 3.0) license.

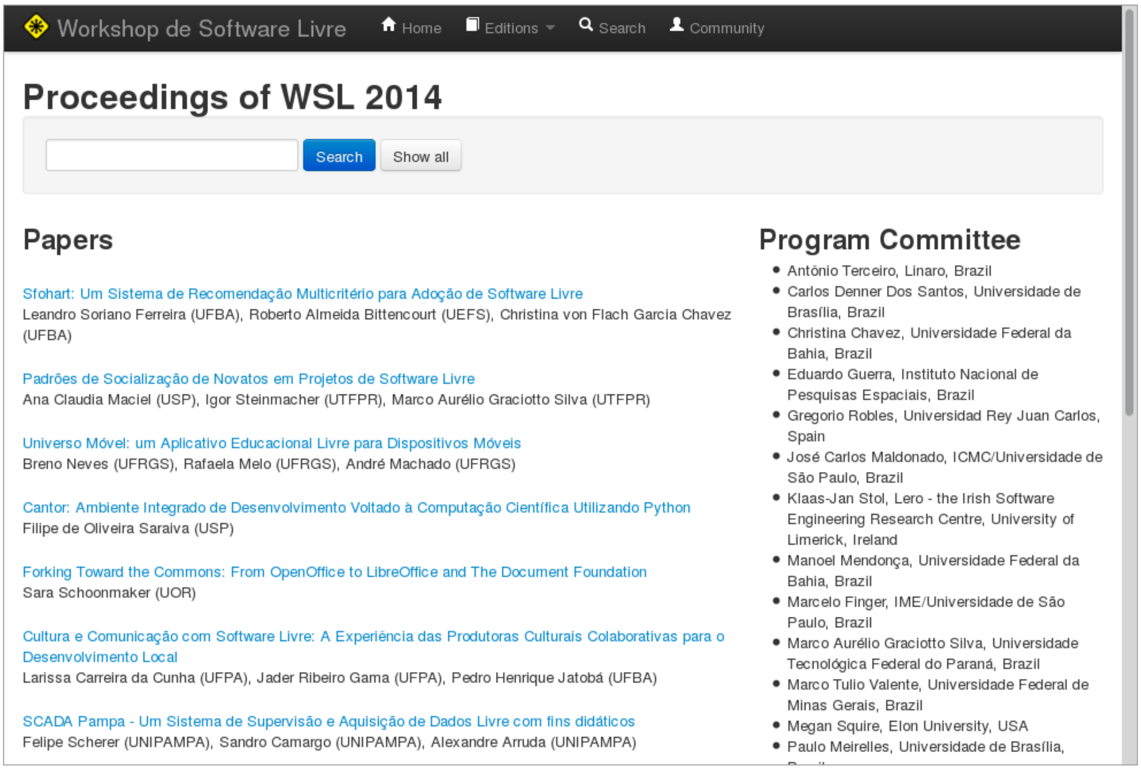
Figure 2. Initial page of the WSL repository

Every page will have a navigation bar at the top containing links to the home page, a dropdown menu with links to each WSL edition, a link to the search page, and a link to the WSL community web page.

The home page also presents information about the repository and the workshop in three different languages: English, Spanish and Portuguese, which are the languages accepted historically for WSL papers. All other text in the repository, except for the papers themselves, are in English only.

[5] pandoc website: http://johnmacfarlane.net/pandoc/

4.2. Proceedings page



Proceedings of WSL 2014

Papers

Sfohart: Um Sistema de Recomendação Multicritério para Adoção de Software Livre
Leandro Soriano Ferreira (UFBA), Roberto Almeida Bittencourt (UEFS), Christina von Flach Garcia Chavez (UFBA)

Padrões de Socialização de Novatos em Projetos de Software Livre
Ana Claudia Maciel (USP), Igor Steinmacher (UTFPR), Marco Aurélio Graciotto Silva (UTFPR)

Universo Móvel: um Aplicativo Educacional Livre para Dispositivos Móveis
Breno Neves (UFRGS), Rafaela Melo (UFRGS), André Machado (UFRGS)

Cantor: Ambiente Integrado de Desenvolvimento Voltado à Computação Científica Utilizando Python
Filipe de Oliveira Saraiva (USP)

Forking Toward the Commons: From OpenOffice to LibreOffice and The Document Foundation
Sara Schoonmaker (UOR)

Cultura e Comunicação com Software Livre: A Experiência das Produtoras Culturais Colaborativas para o Desenvolvimento Local
Larissa Carreira da Cunha (UFPA), Jader Ribeiro Gama (UFPA), Pedro Henrique Jatobá (UFBA)

SCADA Pampa - Um Sistema de Supervisão e Aquisição de Dados Livre com fins didáticos
Felipe Scherer (UNIPAMPA), Sandro Camargo (UNIPAMPA), Alexandre Arruda (UNIPAMPA)

Program Committee

- Antônio Terceiro, Linaro, Brazil
- Carlos Denner Dos Santos, Universidade de Brasília, Brazil
- Christina Chavez, Universidade Federal da Bahia, Brazil
- Eduardo Guerra, Instituto Nacional de Pesquisas Espaciais, Brazil
- Gregorio Robles, Universidad Rey Juan Carlos, Spain
- José Carlos Maldonado, ICMC/Universidade de São Paulo, Brazil
- Klaas-Jan Stol, Lero - the Irish Software Engineering Research Centre, University of Limerick, Ireland
- Manoel Mendonça, Universidade Federal da Bahia, Brazil
- Marcelo Finger, IME/Universidade de São Paulo, Brazil
- Marco Aurélio Graciotto Silva, Universidade Tecnológica Federal do Paraná, Brazil
- Marco Tulio Valente, Universidade Federal de Minas Gerais, Brazil
- Megan Squire, Elon University, USA
- Paulo Meirelles, Universidade de Brasília,

Figure 3. Proceedings of WSL 2014

There is a page for WSL edition. It shows the list of published papers with authors and affiliations, and the program and organizing committee. The search function only applies to papers in that edition.

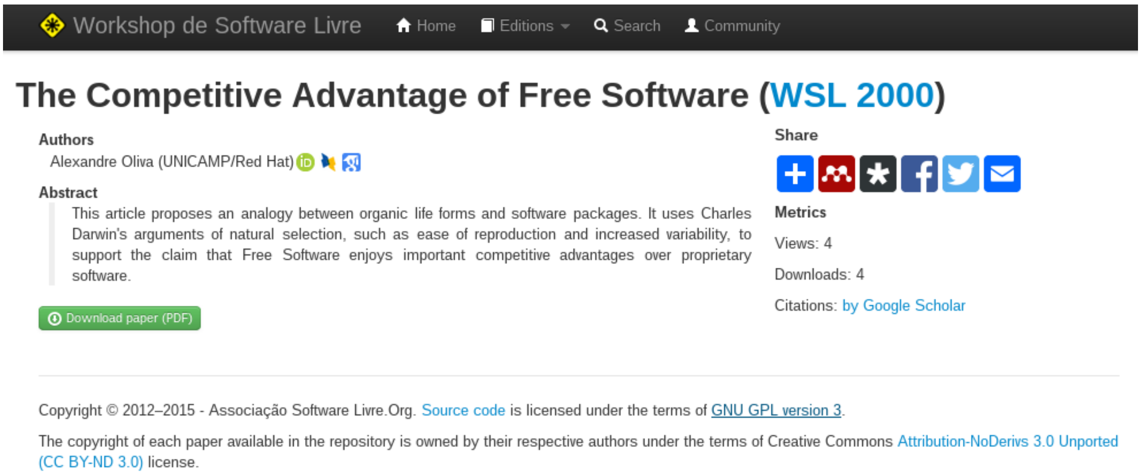
4.3. Paper page

Each paper is presented in its own page, with a link back to corresponding proceedings. The author list features buttons to loop up that author other websites such as ORCID[6], an initiative to provide unique identifiers for researchers; DBLP[7] , the largest bibliographic information repository about computer science publications; and Google Scholar[8], a gratis tool for searching scientific literature, researchers, and to measure the impact of their publications. Extending this list with other services is reasonably easy.

The page presents the abstract and a link to download the paper. For the moment, the repository shows the abstract in the language utilized in the paper. If the paper was written in Portuguese, the abstract will be presented in Portuguese.

In the right-side of the page, there is a side bar with more features and information about the paper. The "Share" buttons uses the AddToAny[9] service, which allows sharing in more than 100 different social networks, like Facebook, Twitter, Diaspora, and services like e-mail, Mendeley, Bibsonomy, and more.

[6] ORCID website: https://www.orcid.org/
[7] DBLP website: http://dblp.uni-trier.de/
[8] Google Scholar website: https://scholar.google.com/
[9] AddToAny website: http://addtoany.com/



The Competitive Advantage of Free Software (WSL 2000)

Authors

Alexandre Oliva (UNICAMP/Red Hat)

Abstract

This article proposes an analogy between organic life forms and software packages. It uses Charles Darwin's arguments of natural selection, such as ease of reproduction and increased variability, to support the claim that Free Software enjoys important competitive advantages over proprietary software.

Metrics

Views: 4

Downloads: 4

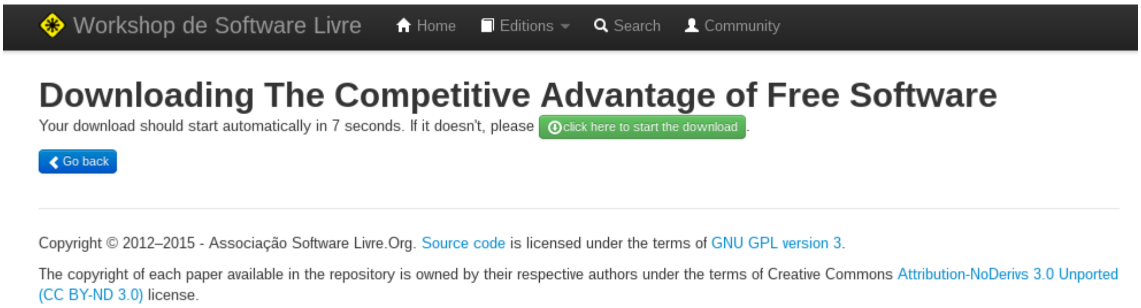
Citations: by Google Scholar

Figure 4. WSL paper information page

Below the "Share" buttons there are some metrics. There is a counter to number of unique visits and downloads. The paper page and the download page have a JavaScript code to send information for a piwik[10] installation maintained by the ASL (*Associação Software Livre*). Piwik is a free software application for web site access analytics.

Lastly, there is a link to Google Scholar in order to show the number of citations of the paper. Google Scholar does not have an API to allow the repository to extract this data and show it directly in the website, so the link to Google Scholar is necessary to use this feature.

4.4. Download page

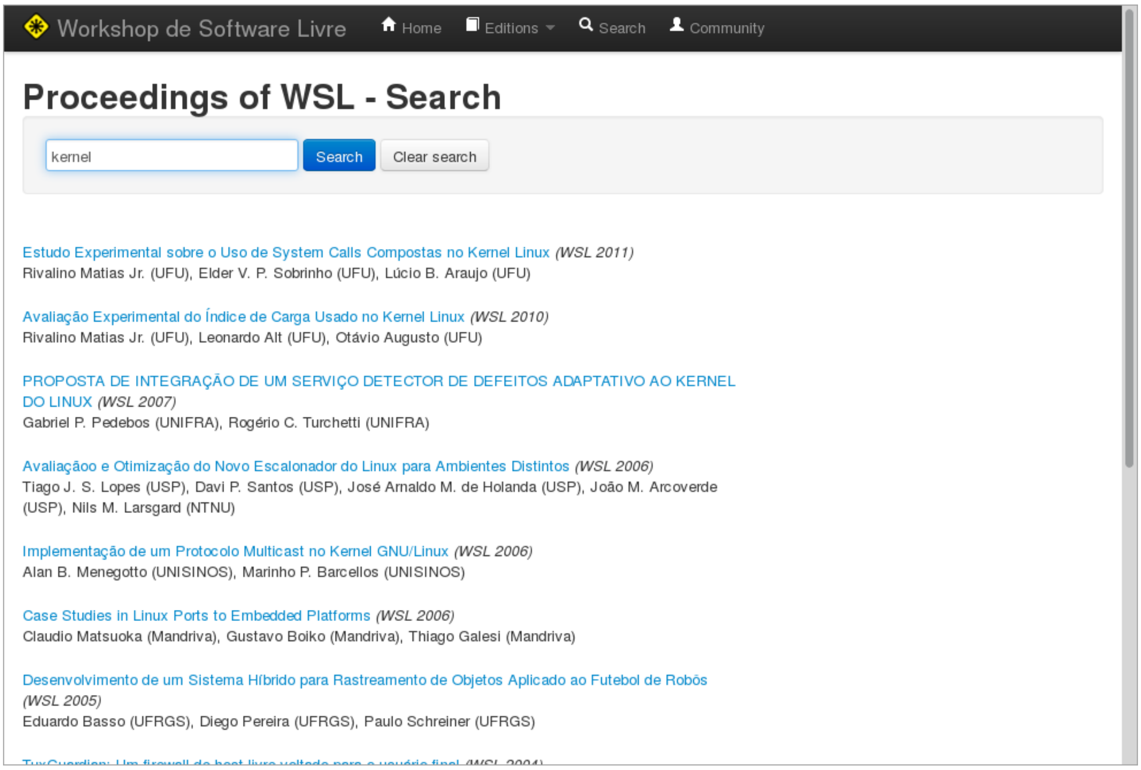


Downloading The Competitive Advantage of Free Software

Your download should start automatically in 7 seconds. If it doesn't, please click here to start the download.

Figure 5. Download page example

This page presents the paper title, a link to download the paper, and a decreasing counter beginning in ten seconds. This counter is there just to allow the download counter JavaScript to access the piwik instance and increment the paper download counter.



Proceedings of WSL - Search

Estudo Experimental sobre o Uso de System Calls Compostas no Kernel Linux *(WSL 2011)*
Rivalino Matias Jr. (UFU), Elder V. P. Sobrinho (UFU), Lúcio B. Araujo (UFU)

Avaliação Experimental do Índice de Carga Usado no Kernel Linux *(WSL 2010)*
Rivalino Matias Jr. (UFU), Leonardo Alt (UFU), Otávio Augusto (UFU)

PROPOSTA DE INTEGRAÇÃO DE UM SERVIÇO DETECTOR DE DEFEITOS ADAPTATIVO AO KERNEL DO LINUX *(WSL 2007)*
Gabriel P. Pedebos (UNIFRA), Rogério C. Turchetti (UNIFRA)

Avaliaçãoo e Otimização do Novo Escalonador do Linux para Ambientes Distintos *(WSL 2006)*
Tiago J. S. Lopes (USP), Davi P. Santos (USP), José Arnaldo M. de Holanda (USP), João M. Arcoverde (USP), Nils M. Larsgard (NTNU)

Implementação de um Protocolo Multicast no Kernel GNU/Linux *(WSL 2006)*
Alan B. Menegotto (UNISINOS), Marinho P. Barcellos (UNISINOS)

Case Studies in Linux Ports to Embedded Platforms *(WSL 2006)*
Claudio Matsuoka (Mandriva), Gustavo Boiko (Mandriva), Thiago Galesi (Mandriva)

Desenvolvimento de um Sistema Híbrido para Rastreamento de Objetos Aplicado ao Futebol de Robôs *(WSL 2005)*
Eduardo Basso (UFRGS), Diego Pereira (UFRGS), Paulo Schreiner (UFRGS)

TuxGuardian: Um firewall de host livre voltado para o usuário final *(WSL 2004)*

Figure 6. Searching for “kernel” on all WSL papers

4.5. Search page

This page provides a search functionality that applies to all papers published in all WSL editions. Queries are applied to text in the title, abstract, and authors names and affiliaton.

The results present the paper title and link, the WSL edition where it was published, authors names and affiliation. It is possible to clear the search result pressing a button “Clear search” in the page.

5. Conclusions

In this paper we described the design and implementation of a free software-based, open access online repository for the whole history of papers published in all past editions of WSL.

The fact that we have all manuscripts and metadata in a single repository opens up the possibility for the easy execution of secondary studies such as systematic reviews and mapping studies. We look forward to seeing such studies across the WSL history.

Other possibilities for future work include improvements and development of new features in the repository itself. For example, we could link the authors’ Lattes curriculum, (academic researchers CV platform maintained by the Brazilian government), or the authors’ profiles on BDBComp (Brazilian Computer Society bibliographic service); we can also provide easy tools to measure the number of shares in social networks for each paper.

[10] piwik website: http://piwik.org/

A high priority future endeavor, after the repository has been launched, is submitting the proceedings to conventional academic publication services in order to increase the formal scientific value of the WSL in academic community. We are trying to get an ISSN (International Standard Serial Number) identifier for the conference and a DOI (Digital Objects Identifier) for each paper. After that, the authors will register the proceedings in important computer science bibliographic databases such as DBLP, BDBComp, DOAJ (Directory of Open Access Journals), Google Scholar, and others. Each of these databases has their own XML file format for receiving information about proceedings and papers. We will be able to generate these file formats automatically from the metadata files, in the same way as we generate the HTML for the repository web interface. Automating the process will make it easier for future WSL organizers to keep doing it for future editions.

Any member of the WSL community is welcome to get in touch with the authors to collaborate on any of the above ideas, or to propose and implement new ones.

Last, but not least, we have made the source code of the repository free software so that other publications can benefit from it. We will be glad to drive the process of generalizing all the code we have written to support use cases different than ours. Potential users can contact us using the collaboration tools on GitLab (preferred) or directly.

Acknowledgments

The authors are grateful to all the people involved with the *Workshop de Software Livre* since its beginning. Authors, organizers, reviewers, committee members, and chairs have contributed throughout the years to build a relevant academic conference about free software in several aspects. This open access repository is a tribute to all that work.

We would like specifically to thank Rodrigo Reis for providing the proceeding from WSL 2005; Adenauer Yamin for WSL 2006; Marinho Barcellos for WSL 2007; Patrícia Kayser for WSL 2008; Andrea Charão for WSL 2009; Cristiano Costa WSL 2010; Junior Goergen for the work of scanning the proceedings of WSL 2000 and 2002; Wille Marcel for reviewing the Spanish translation of the web interface; and Cristina Atz for the help with some verifications against the printed versions.

Finally, thanks to *Associação Software Livre* for making FISL happen year after year and for making it possible to always host WSL together with FISL.

References

BOAI (2012). Ten years on from the Budapest Open Access Initiative: setting the default to open.

Chan, L., Cuplinskas, D. and Eisen, M.et al. (2002). Budapest Open Access Initiative.

Laakso, M., Welling, P. and Bukvova, H.et al. (2011). The development of open access journal publishing from 1993 to 2009. *PloS one*, v. 6, n. 6, p. e20961.

Muir, S. P., Leggott, M. and Willinsky, J. (2005). Open journal systems: An example of open source software for journal management and publishing. *Library hi tech*, v. 23, n. 4, p. 504–519.

Swan, A. (2012). *Policy guidelines for the development and promotion of open access*. Tradução. UNESCO.