Better render engine for web page archiving

Egg, September 21, 2022: *The import of HTML content (web pages) for the ArchivistaBox was previously managed via PDF files from the URL links. Now HTML content can be implemented directly via the Firefox engine or the SingleFile add-on. Version 2022/IX also brings a whole host of innovations, more about this at the end of the article.*

Previous import via LibreOffice

To show the “weaknesses” in the previous import, this post uses the last ArchivistaBox blog entry: **2022/VIII and new Prices.** To save a (this) page as HTML file, it has to be saved as HTML page in the web browser. Then it can be moved to the Office folder of the ArchivistaBox.

Version 2022/VIII mit viel Liebe zu Sicherheit

Nach dem Motto ‘Wer rastet, der rostet’ werden Software-Lösungen gegenwärtig in einer Geschwindigkeit weiterentwickelt, dass dabei oftmals die Liebe zum Detail, vor allem aber die Aspekte der Sicherheit zu kurz kommen. Aus diesem Grund wurde die ArchivistaBox die letzten beiden Monate quasi eingefroren. Anstelle von neue Funktionen, wurde die Software dort verbessert, wo mehr Finetuning viel Sinn ergibt. Zudem wurde der Linux-Unterbau mit den neusten stabilen Versionen ausgestattet.

Gemäss **https://stack.watch/product/linux/linux-kernel** gibt es im Linux-Kernel bis zum September 2022 bereits 178 Bugs, die beseitigt werden mussten. Wer jetzt meinte, mit Windows würde dies nicht passieren, dem darf entgegnet werden, dass bei Windows 11 gemäss **https://stack.watch/product/microsoft/windows-11** für 2022 bislang 333 Bugs gemeldet sind. Nicht enthalten sind dabei Bugs auf tieferer Stufe (BIOS). Diese sogenannten UEFI-Bugs sind weit gefährlicher, weil bei Befall ein Austausch bzw. Neuinstallation der Software (Linux/Windows) keine Abhilfe schafft. Der Bug liegt tiefer. Stellvertretend für viele ein **Beitrag von 2.2.22 (das Datum ist Programm -;) von heise.de.**

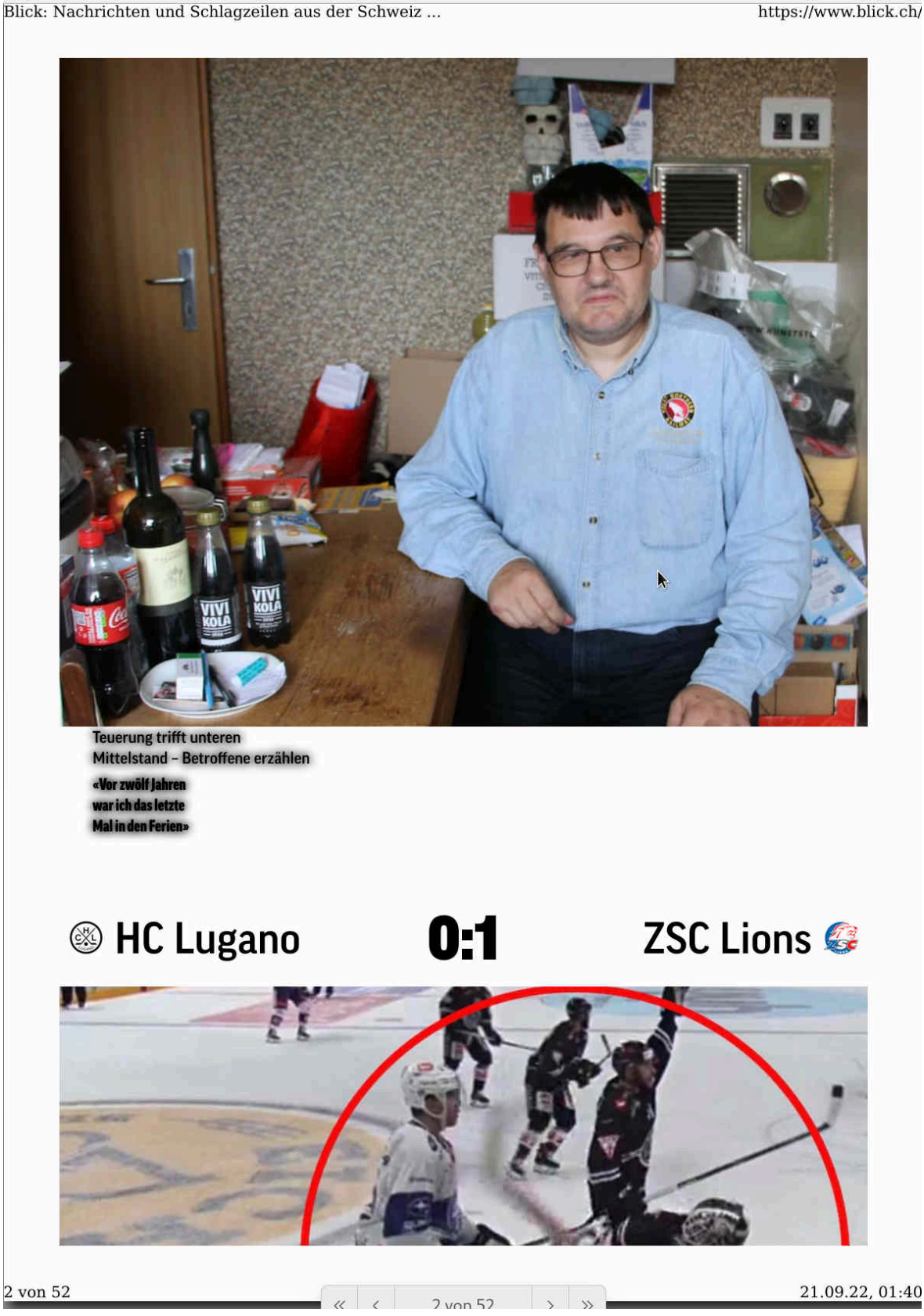
Daher an dieser Stelle die dringende Empfehlung. Bitte Update auf 2022/XIII beziehen und einspielen. Mit WebConfig ist der Vorgang sehr einfach.

During processing LibreOffice is called, a PDF file is created from the HTML file and this in turn is used for the import. This captures basic HTML elements, but when it comes to

something more design (which is now the case with almost all homepages), then the HTML import fails.

Process HTML files as PDF files

In almost all modern browsers it is possible to create a PDF file directly when printing. Common pages (our example is one of them) of a webpage can be captured in satisfying quality with it.



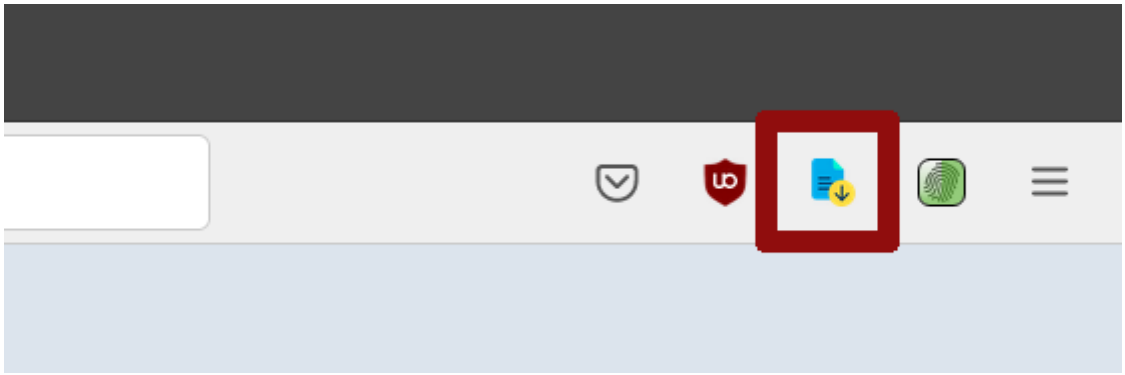
It becomes more difficult when content providers use the latest HTML tricks. The homepage of blick.ch can be cited as an example. The PDF files created (e.g. from Firefox) are unfortunately not able to “conjure up” anything like what was originally

present on the homepage.

HTML files with Firefox AddOn SingleFile

If you want to archive web pages with an exact layout, you are first confronted with the problem that an HTML page consists of many small puzzle pieces. If you simply save the HTML file with ‘Save as’, you will find a lot of other files in the file manager besides the HTML file.

It is therefore not enough to simply transfer the HTML file to the ArchivistaBox, because this means that the images and layout templates (CSS files) are missing. Likewise, HTML files often and readily contain external script files, which are absolutely necessary for correct display of the content.

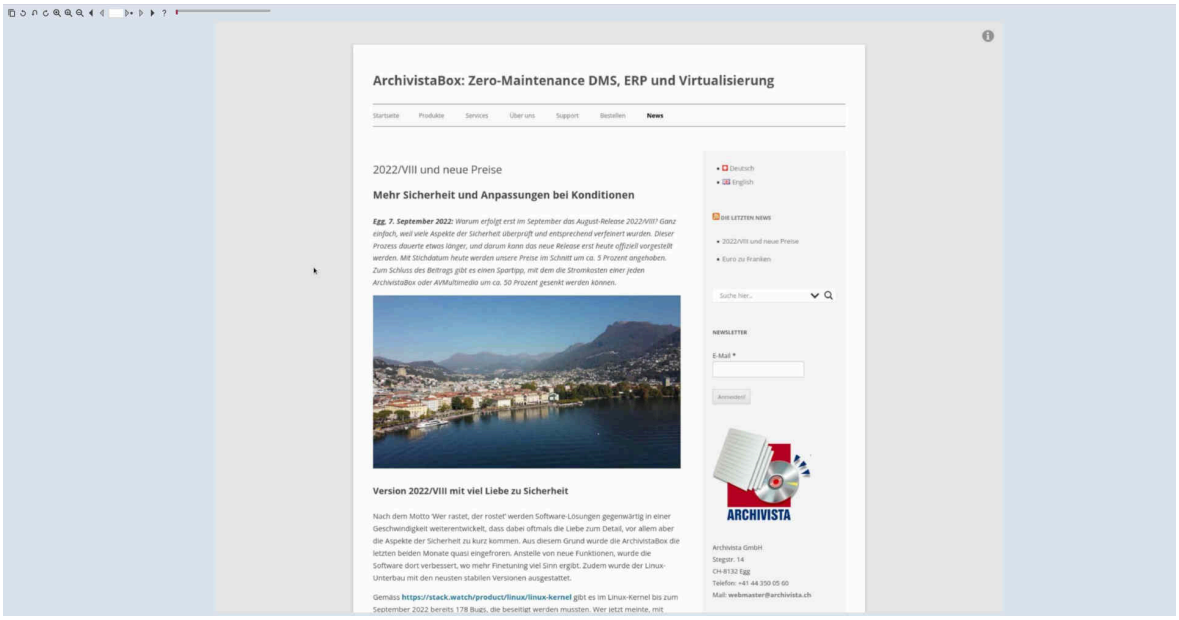


Step 1: Export the desired page with SingleFile

This is where the Firefox add-on ‘SingleFile’ comes in. This is included on the ArchivistaBox from version 2022/IX together with updated Firefox. The button (icon) is located at the top right of the browser. One click on it is sufficient for all external HTML elements to be “freighted” into the HTML file created by SingleFile.

Step 2: Import as HTML file to ArchivistaBox

This file can then be conveniently transferred via the file upload or the ArchivistaBox share folder (office directory).



In contrast to previous versions, HTML files are now processed with Firefox itself in the so-called headless mode. The browser is opened in the background during processing and a screen copy is created at the same time using the HTML file. In addition, the text is extracted directly from the original and transferred to the ArchivistaBox. Analogous to the above example, here is the blick.ch page again:

Especially with complex layouts (e.g. white text in images), significantly better results can be achieved in this way than would be the case if either the text recognition were “fired up” or the text were read out via a generated PDF file.

Conclusion: Perfect HTML imports as of version 2022/IX

The results are impressive. Even complete layouts are processed perfectly, the generated pages in the ArchivistaBox offer almost 100% accuracy of fit to the originals.

By the way: All those who object that it would be easier to display imported HTML files directly as HTML files (e.g. in a new tab) in a browser window, should be aware that clicking on ‘File’ for the respective document in the ArchivistaBox will produce exactly that.

Only, who can guarantee that in five or ten years time the respective browsers will be able to display the content in its original form, if (after approx. 30 years of web technology) the printing of more complex HTML pages still does not work satisfactorily?

Only “virtual scanning” results in long-term archiving

For this reason alone, virtual “scanning” of content is essential, especially for HTML content. With “rasterized” copies of the content, the ArchivistaBox has been offering long-term archiving of the entered data for almost 25 years now, which is not otherwise available on the market. Representative of many is the concept for the long-term archiving of the university.

			Planning
PDF/A-1	1	• Offener Standard • Weltweit unterstützt und weit verbreiteter ISO Standard • Basiert auf PDF Version 1.4 • Alle Bilder, Graphiken, Schriften müssen eingebettet sein. • Farben müssen gerätunabhängig sein • Transparente Elemente, JavaScript (ausführbarer Code), Multimedia sind nicht erlaubt • Darf nicht passwortgeschützt oder verschlüsselt sein • Darf keine Links nach aussen enthalten • Darf keine audio, video oder 3D Daten enthalten. • XMP für Metadaten (Autor*in, Thema, Inhalt, Erstellungsdatum, etc.) • PDF/A-1a: Entspricht vollumfänglich dem PDF/A-1b Standard und hat Merkmale, die für die barrierefreie Zugänglichkeit von Inhalten und ihre Darstellung auf Mobilgeräten wichtig sind (sog. «tagged PDF»). • PDF/A-1b: Gewährleistet die langfristige Erhaltung des Erscheinungsbilds eines Dokuments	• Wird als Abgabeformat empfohlen

It lists which formats are suitable for long-term archiving. Interestingly, the HTML format is not listed. Instead, either the Tiff format or primarily PDF/A is recommended. It is astonishing that after 30 years of WWW, HTML is not even mentioned as a source in this guide.

It should also be noted that although embedded JavaScript code is “permitted” in PDF/A files, all audio and video content must explicitly not be transferred to long-term archiving in accordance with this concept.

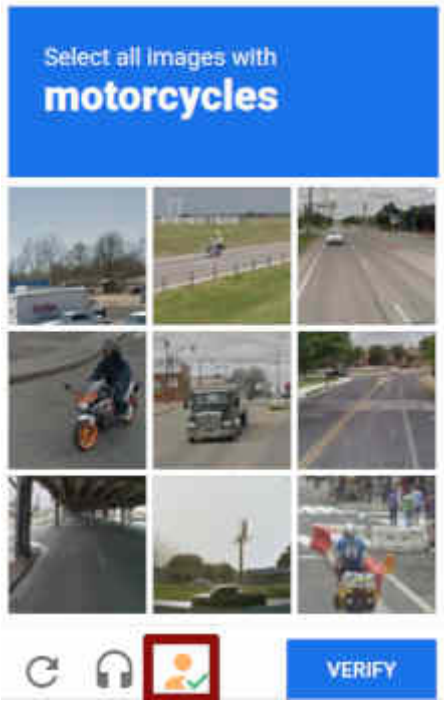
Other new features in version 2022/IX

The following innovations can be found on both AVMultimedia and ArchivistaBox. As already mentioned above, version 2022/IX is delivered with updated Firefox (currently version 105). The AddOns have also been updated and cleaned up. No longer on board is Hide-My-IP, as the add-on is no longer “maintained”. With the integration of ProtonVPN, there is basically a worthy alternative available.

Buster: Captcha-Solver-for-Humans

Annoying in the last months is the captcha mania everywhere. And even more annoying is that once again the monopoly ‘reCaptcha’ of the search giant is spreading almost epidemically. This may be good for the giant (every click generates user data), but it is currently a plague. Even if the add-on ‘Buster: Captcha-Solver-for-Humans’ does not

eliminate all reCaptchas, in many cases the click offers a remedy, tedious image “orgies” can be avoided.



Technologically, the solution works by sending the audio stream to the Buster server. This “solves” the puzzle with speech recognition. Since reCaptcha is “fed” with new nonsensical examples everywhere, it can unfortunately sometimes happen that the speech recognition does not lead to the desired result. Clicking on images is still possible.

UBlock and CanvasBlocker

Also included on the ArchivistaBox is UBlock, in order to be able to surf the Internet as carefree as possible. Two or three discreet addons on a search mask, no one would have anything to complain about. The search giant started out with exactly that in mind. Only, meanwhile surfing the net turns into a gauntlet, so much advertising is displayed that UBlock at least provides some relief.

New or as an alternative to the previous fingerprinting is CanvasBlocker included. Since most homepages are now “overloaded” with cookies, which can be used to seamlessly “track” surfing behavior with unique keys, CanvasBlocker basically generates random garbage to make tracking (at least) more difficult.

Restore Firefox startup folder on restart.

Both AddOns work in such a way that there is a learning curve, i.e. the tools learn based on browsing behavior. So that the “training effect” is not lost after each restart, the .mozilla folder under /home/archivista can now be retained on restart, provided it has previously been copied to /home/archivista/data/mozilla.

The above procedure is also useful if you want to activate your own add-ons. When creating the .mozilla folder, it is of course important that cached files are eliminated beforehand.

ProtonVPN: Surfing across national borders

In itself, the net would be open for everyone and everywhere at any time. However, so-called geo-locks are being activated more and more frequently. For example, if you want to watch a movie on Arte.tv, you will find (if you are in Switzerland) the terse information that the content is not available in this language region.

With ProtonVPN, this can be “overridden”. It is important to mention that it is not possible at this point to discuss whether virtual private networks are legal or illegal in all countries or in which states. Moreover, the use of ProtonVPN requires an account with the provider and only a few locations (Netherlands and USA) are available in the free version. If you want more, you have to buy a subscription (costs currently between 5 and 15 francs/euro for ten devices).

The console version is currently available from ArchivistaBox or AVMultimedia. The account can be set up with ‘protonvpn init’. If you want to have this process activated beyond the reboot, you must activate the files or folders in each case. The following script may serve as assistance, how this can be accomplished:

```
#!/bin/bash
pin="/home/archivista/data/protonvpn"
cp -pf $pin/update-resolv-conf /etc/openvpn
cp -pf $pin/resolv.conf /etc
cp -rpf $pin/.pvpn-cli /home/archivista
```

Automation on the desktop with xmacro

Who does not know the problem. Again and again the same processes occur on the desktop. With the package xmacro these processes can be automated. For this purpose the programs ‘xmacrorec2’ and ‘xmacroplay’ are available.

Addendum from 22.9.22: *On* ***linuxnews.de a news about version 2022/IV has been published.*** *linuxnews.de in particular is a good illustration of what the new HTML import for the ArchivistaBox brings. The page* ***transferred to the ArchivistaBox offers an accurate view,*** *the* ***PDF printed from Firefox on the other hand looks quite “old-fashioned”*** *— and moreover requires five times the memory.*