Bessere Render-Engine für Webseiten-Archivierung

Egg, 21. September 2022: *Der Import von HTML-Inhalten (Web-Seiten) für die ArchivistaBox wurde bisher über PDF-Dateien ab den URL-Links bewerkstelligt. Neu können HTML-Inhalte direkt über die Firefox-Engine bzw. das AddOn SingleFile realisiert werden. Überdies bringt Version 2022/IX einen ganzen Strauss an Neuerungen, dazu mehr am Ende des Beitrages.*

Bisheriger Import über LibreOffice

Um die «Schwächen» beim bisherigen Import aufzuzeigen, wird in diesem Beitrag der letzte Blog-Eintrag der ArchivistaBox verwendet: **2022/VIII und neue Preise.** Um eine (diese) Seite als HTML-Datei zu sichern, ist sie im Web-Browser als HTML-Seite zu speichern. Danach kann sie in den Office-Ordner der ArchivistaBox verschoben werden.

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.**

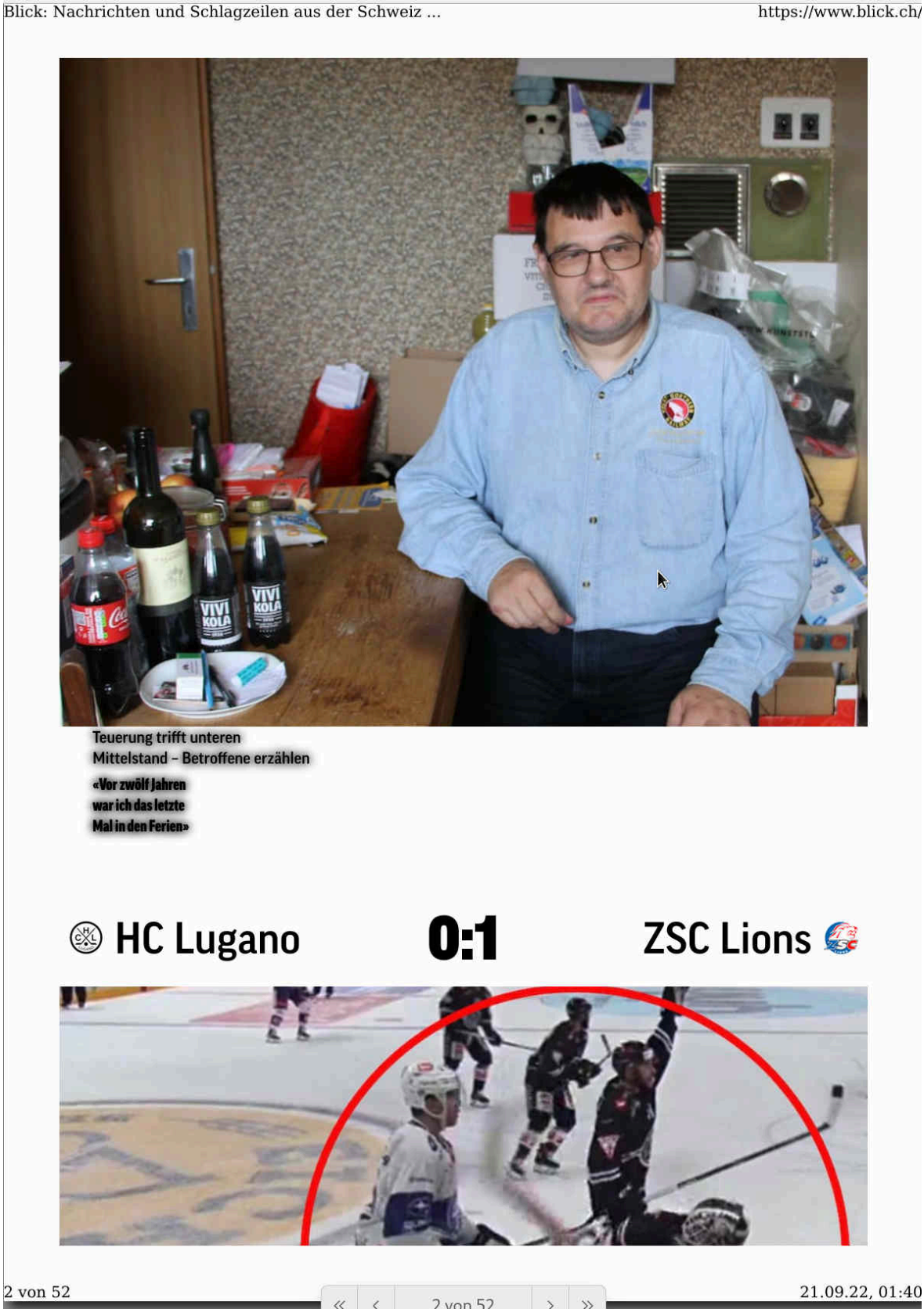
Daher an dieser Stelle die dringende Empfehlung. Bitte Update auf 2022/XIII beziehen und einspielen. Mit WebConfig ist der Vorgang sehr einfach.

Bei der Verarbeitung wird LibreOffice aufgerufen, aus der HTML-Datei entsteht eine

PDF-Datei und diese wiederum wird für den Import verwendet. Dabei werden zwar grundlegende HTML-Elemente erfasst, doch geht es um etwas mehr Gestaltung (was mittlerweile bei fast allen Homepages der Fall ist), dann versagt der HTML-Import.

HTML-Dateien als PDF-Dateien verarbeiten

In fast allen modernen Browsern ist es möglich, beim Drucken direkt eine PDF-Datei zu erstellen. Übliche Seiten (dazu zählt auch unser Beispiel) einer Webpage können damit in zufriedenstellender Qualität erfasst werden.



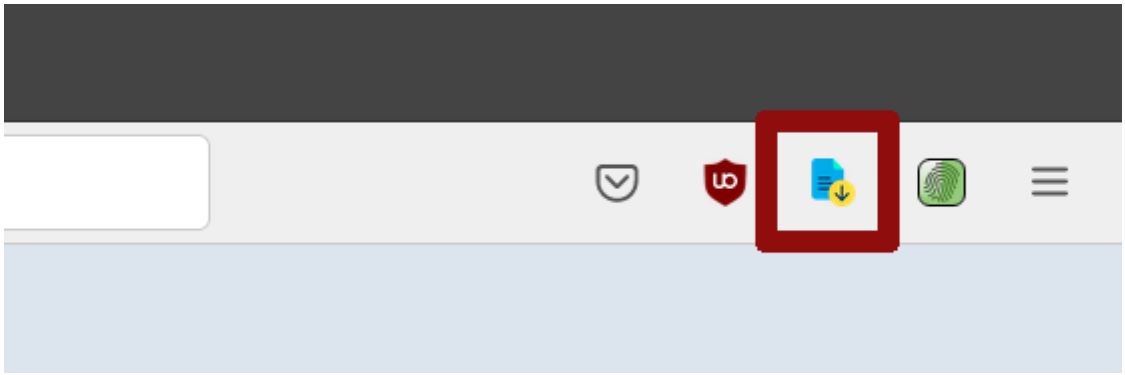
Schwieriger wird es, wenn die Inhaltsanbieter die letzten HTML-Kniffs verwenden. Als Beispiel sei hier die Homepage von blick.ch angeführt. Die (z.B. ab Firefox) erstellen PDF-Dateien vermögen leider nicht in Ansätzen das «hervorzuzaubern», was originär

auf der Homepage vorhanden war.

HTML-Dateien mit Firefox-Addon SingleFile

Wer Webseiten layoutgenau archivieren möchte, sieht sich zunächst einmal mit dem Problem konfrontiert, dass eine HTML-Seite aus vielen kleinen Puzzleteilen besteht. Wer folglich einfach die HTML-Datei mit ‹Speichern unter› sichert, findet im Dateimanager neben der eigentlichen HTML-Datei meist eine Unzahl weiterer Dateien vor.

Es reicht folglich nicht, einfach die HTML-Datei zur ArchivistaBox zu übertragen, weil damit die Bilder und Layoutvorlagen (CSS-Dateien) fehlen. Ebenso enthalten HTML-Dateien oft und gerne externe Skript-Dateien, die für ein korrektes Anzeigen der Inhalte zwingend notwendig sind.

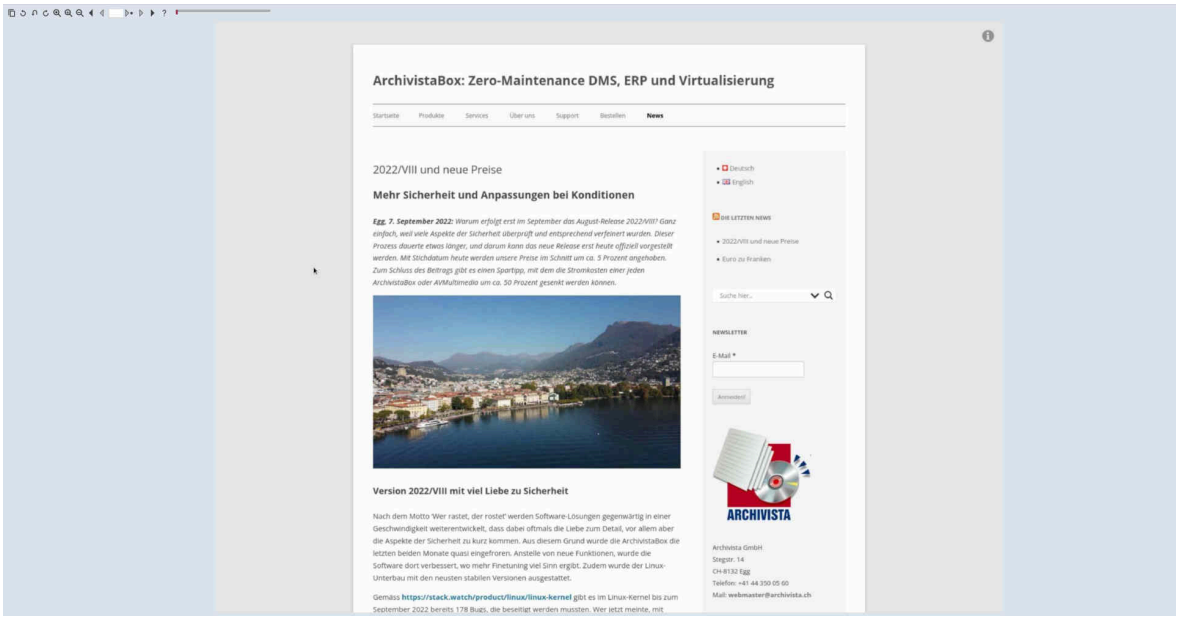


Schritt 1: Export der gewünschten Seite mit SingleFile

An dieser Stelle setzt das Firefox AddOn ‹SingleFile› ein. Dieses ist ab Version 2022/IX zusammen mit aktualisiertem Firefox auf der ArchivistaBox enthalten. Der Button-Knopf (Icon) befindet sich oben rechts im Browser. Ein Klick darauf genügt, dass alle externen HTML-Elemente in die von SingleFile erstellen HTML-Datei «verfrachtet» werden.

Schritt 2: Import als HTML-Datei zur ArchivistaBox

Diese Datei kann danach bequem über den Datei-Upload bzw. den ArchivistaBox-Freigabe-Ordner (office-Verzeichnis) übertragen werden.



Im Unterschied zu früheren Versionen werden HTML-Dateien dabei neu mit Firefox selber im sogenannten Headless-Modus verarbeitet. Der Browser wird im Hintergrund bei der Verarbeitung geöffnet und gleichzeitig wird anhand der HTML-Datei eine Bildschirmkopie erstellt. Zusätzlich wird der Text direkt aus dem Original extrahiert und in die ArchivistaBox überführt. Analog zum obigen Beispiel hier auch nochmals die blick.ch-Seite:

Gerade bei aufwändigen Layouts (z.B. weisse Schrift in Bildern) können so deutlich bessere Resultate erzielt werden, als dies der Fall wäre, wenn entweder die Texterkennung «angeworfen» würde oder der Text über eine generierte PDF-Datei ausgelesen würde.

Fazit: Perfekte HTML-Importe ab Version 2022/IX

Die Resultate können sich sehen lassen. Selbst komplette Layouts werden perfekt verarbeitet, die generierten Seiten in der ArchivistaBox bieten nahezu 100% Passgenauigkeit zu den Originalen.

Übrigens: All jenen, die an einwenden, es wäre doch einfacher, eingelesene HTML-Dateien jeweils direkt als HTML-Datei (z.B. in einem neuen Tab) in einem Browser-Fenster darzustellen, sei gesagt: Wer in der ArchivistaBox beim jeweiligen Dokument auf ‹Datei› klickt, erhält genau dies.

Nur, wer garantiert, dass in fünf oder zehn Jahren die jeweiligen Browser die Inhalte in der ursprünglichen Form darstellen können, wenn (nach ca. 30 Jahren Web-Technologie) das Drucken von komplexeren HTML-Seiten noch immer nicht befriedigend funktioniert?

Erst «virtuelles Scannen» ergibt Langzeitarchivierung

Alleine aus diesem Grund ist das virtuelle «Scannen» der Inhalte essentiell, gerade bei HTML-Inhalten. Die ArchivistaBox bietet mit «gerasterten» Kopien der Inhalte seit nunmehr fast 25 Jahren eine Langfristigkeit bei den eingepflegten Daten an, die es sonst auf dem Markt so nicht gibt. Stellvertretend für viele sei auf das Konzept für die Langzeitarchivierung der Universität verwiesen.

			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

Darin wird aufgelistet, welche Formate sich für die Langzeitarchivierung eignen. Interessanterweise findet sich das HTML-Format darin gerade nicht. Vielmehr wird entweder das Tiff-Format oder primär PDF/A empfohlen. Schon erstaunlich, nach mittlerweile ca. 30 Jahren WWW wird in diesem Leitfaden HTML gar nicht erst als Quelle angeführt.

Ebenso darf angeführt werden, dass zwar eingebetteter JavaScript-Code in den PDF/A-Dateien «zulässig» ist, alle Audio- und Video-Inhalte dürfen explizit nicht in die Langzeitarchivierung gemäss diesem Konzept überführt werden.

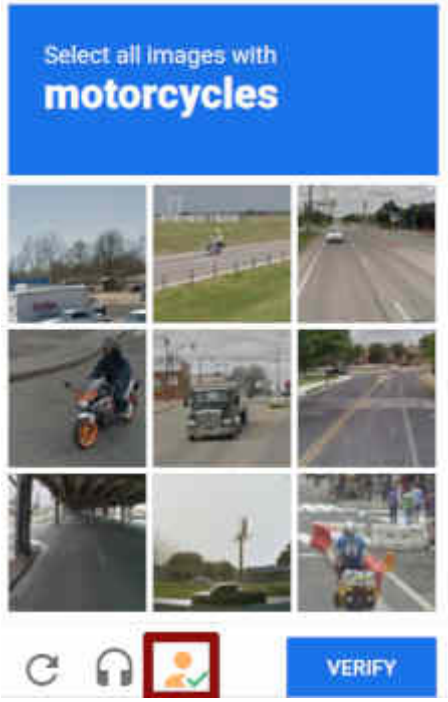
Weitere Neuerungen in Version 2022/IX

Die nachfolgenden Neuerungen finden sich sowohl auf AVMultimedia wie auf der ArchivistaBox. Wie bereits oben angeführt, die Version 2022/IX wird mit aktualisiertem Firefox (aktuell Version 105) ausgeliefert. Ebenso aktualisiert bzw. bereinigt wurden die AddOns. Nicht mehr an Bord ist Hide-My-IP, da das AddOn nicht mehr «gepflegt» wird. Mit der Integration von ProtonVPN steht ganz grundsätzlich eine würdige Alternative zur Verfügung.

Buster: Captcha-Solver-for-Humans

Lästig in den letzten Monaten gestaltet sich die Captcha-Manie allenthalben. Und noch lästiger dabei ist, dass einmal mehr das Monopol ‹reCaptcha› des Suchriesen sich fast

seuchenmässig ausbreitet. Das mag gut für den Giganten sein (ein jeder Klick generiert Nutzerdaten), es ist aber aktuell eine Plage. Auch wenn das AddOn ‹Buster: Captcha-Solver-for-Humans› nicht alle reCaptchas eliminiert, in vielen Fällen bietet der Klick Abhilfe, mühsamste Bilder-«Orgien» können vermieden werden.



Technologisch betrachtet arbeitet die Lösung so, dass der Audio-Stream an den Buster-Server gesendet wird. Dieser «löst» das Rätsel mit Spracherkennung. Da reCaptcha allenthalben mit neuen unsinnigen Beispielen «gefüttert» wird, kann es leider manchmal vorkommen, dass die Spracherkennung nicht zum gewünschten Resultat führt. Bilder-Klicken ist dann ja noch immer möglich.

UBlock und CanvasBlocker

Weiter auf der ArchivistaBox enthalten ist UBlock, um möglichst sorgenfrei im Internet surfen zu können. Zwei drei diskrete Adds auf einer Suchmaske, dagegen hätte wohl niemand etwas zu meckern. Der Suchriese ist ja genau damit angetreten. Nur, mittlerweile gestaltet sich das Surfen im Netz zum Spiessrutenlauf, es wird derart viel Werbung eingeblendet, dass UBlock hier zumindest eine gewisse Abhilfe ergibt.

Neu bzw. als Alternative zum bisherigen Fingerprinting ist CanvasBlocker enthalten. Da mittlerweile der grösste Teil der Hompeages mit Cookies «überfrachtet» ist, anhand derer eindeutiger Keys das Surf-Verhalten nahtlos «getrackt» werden kann, generiert CanvasBlocker im Prinzip zufälligen Müll, damit das Tracking (zumindest) erschwert wird.

Firefox-Startordner beim Neustart wiederherstellen

Beide AddOns arbeiten so, dass es eine Lernkurve gibt, d.h. die Tools lernen aufgrund des Surf-Verhaltens. Damit der «Trainingseffekt» nicht nach jedem Neustart verlorengeht, kann neu der .mozilla-Ordner unter /home/archivista beim Neustart beibehalten werden, sofern er zuvor nach /home/archivista/data/mozilla kopiert wurde.

Das obige Vorgehen bietet sich auch dann an, wenn eigene AddOns aktiviert werden sollen. Wichtig beim Anlegen des .mozilla-Ordners ist natürlich, dass gecachte Dateien zuvor eliminiert werden.

ProtonVPN: Surfen über Landesgrenzen

An sich wäre das Netz ja offen für alle und überall zu jeder Zeit. Immer häufiger werden jedoch sogenannte Geo-Sperren aktiviert. Wer z.B. einen Film bei Arte.tv ansehen möchte, findet (falls in der Schweiz unterwegs) den lapidaren Hinweis, dass die Inhalte in dieser Sprachregion nicht zur Verfügung stünden.

Mit ProtonVPN lässt sich dies «aushebeln». Wichtig zu erwähnen ist, dass an dieser Stelle keine Abhandlung erfolgen kann, ob virtuelle private Netzwerke in allen Ländern bzw. in welchen Staaten legal bzw. illegal sind. Überdies erfordert der Einsatz von ProtonVPN ein Konto beim Anbieter und in der freien Version sind nur einige wenige Standort (Niederlande und USA) verfügbar. Wer mehr möchte, muss ein Abo lösen (Kosten aktuell zwischen 5 bis 15 Franken/Euro für zehn Geräte).

Bei der ArchivistaBox bzw. AVMultimedia ist aktuell die Konsolen-Version verfügbar. Mit ‹protonvpn init› kann das Konto eingerichtet werden. Wer diesen Vorgang über den Neustart hinaus aktiviert haben möchte, muss die Dateien bzw. Ordner jeweils aktivieren. Das nachfolgende Skript möge als Hilfestellung dienen, wie dies bewerkstelligt werden kann:

```
#!/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
```

Automatisierung auf dem Desktop mit xmacro

Wer kennt nicht das Problem. Immer wieder fallen auf dem Desktop die gleichen Abläufe an. Mit dem Paket xmacro können diese Abläufe automatisiert werden. Dazu stehen neu die Programme ‹xmacrorec2› und ‹xmacroplay› zur Verfügung.

Nachtrag vom 22.9.22: *Auf* ***linuxnews.de ist eine News zur Version 2022/IV erschienen.*** *Gerade auch an linuxnews.de lässt sich gut veranschaulichen, was der neue HTML-Import für die ArchivistaBox bringt. Die zur* ***ArchivistaBox übertragene Seite bietet eine passgenaue Ansicht,*** *das aus* ***Firefox heraus gedruckte PDF wirkt dagegen recht «altbacken»*** *— und benötigt überdies den fünfachen Speicherbedarf.*