# WARC hands on workshop

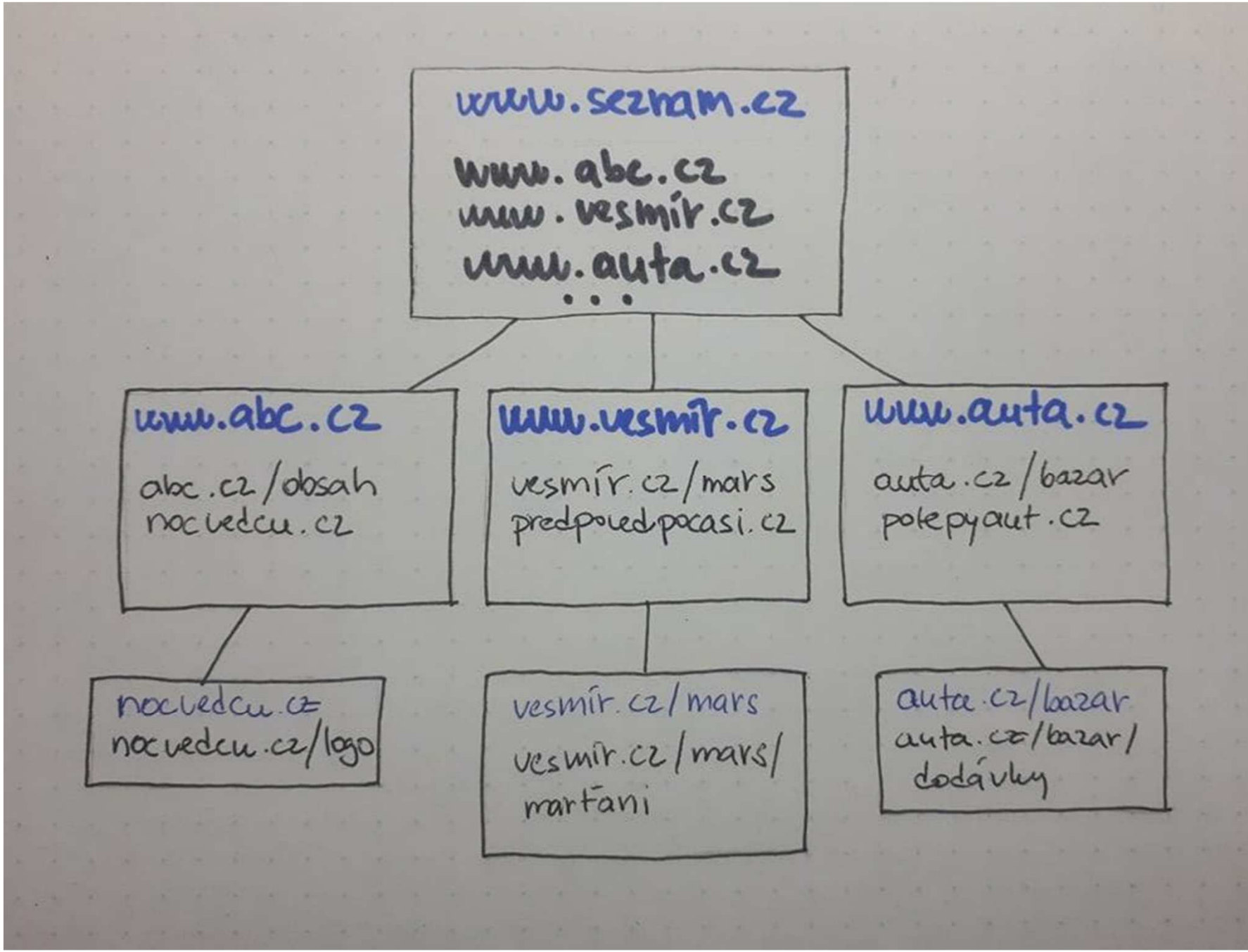
*Praktická ukázka práce s WARC soubory*

# *Co nás čeká:*

- Trocha teorie: WARC, Heritrix, index, sklízení webu.
- Vytvoření vlastního kontejneru pomocí nástroje Webrecorder.
- Prozkoumání vytvořeného kontejneru.
- (Vytvoření vlastního indexu.)

# *webový sklízeč*

- sklízeč neboli crawler je internetový robot
- crawler navštěvuje URL adresy ze zadaného seznamu adres (1 400 000 aktuálně)
- různé hloubky sklízení (kolik odkazů navštíví na jedné stránce)
- výsledkem sklízení je index


www.seznam.cz
www.abc.cz
www.vesmir.cz
www.auta.cz
...
www.abc.cz
abc.cz/obsah
nocvedcu.cz
www.vesmir.cz
vesmir.cz/mars
predpovedpocasi.cz
www.auta.cz
auta.cz/bazar
pokepyaut.cz
nocvedcu.cz
nocvedcu.cz/logo
vesmir.cz/mars
vesmir.cz/mars/
martani
auta.cz/bazar
auta.cz/bazar/
dodavky

# *Heritrix*

- open-source nástroj společnosti Internet Archive
- současná verze 3.4
- požadavky: min 4 GB RAM paměti, dostatek místa k vytvoření WARC složky, co nejméně konkurenčních procesů
- jazyk: Java, Javascript

# *Kdo používá Heritrix?*

- Internet Archive (USA)
- The British Library (Velká Británie)
- The Library of Congress (USA)
- California Digital Library, Web archiving service (USA)
- BNCF, Biblioteca Nazionale Centrala Firenze (Itálie)
- Smithsonian Institution Archives (USA)
- Netarchive.dk (Dánsko)
- National and University Library of Iceland (Island)
- French National Library (Francie)
- Austrian National Library (Rakousko)
- National Library of Catalonia (Španělsko)

# *index*

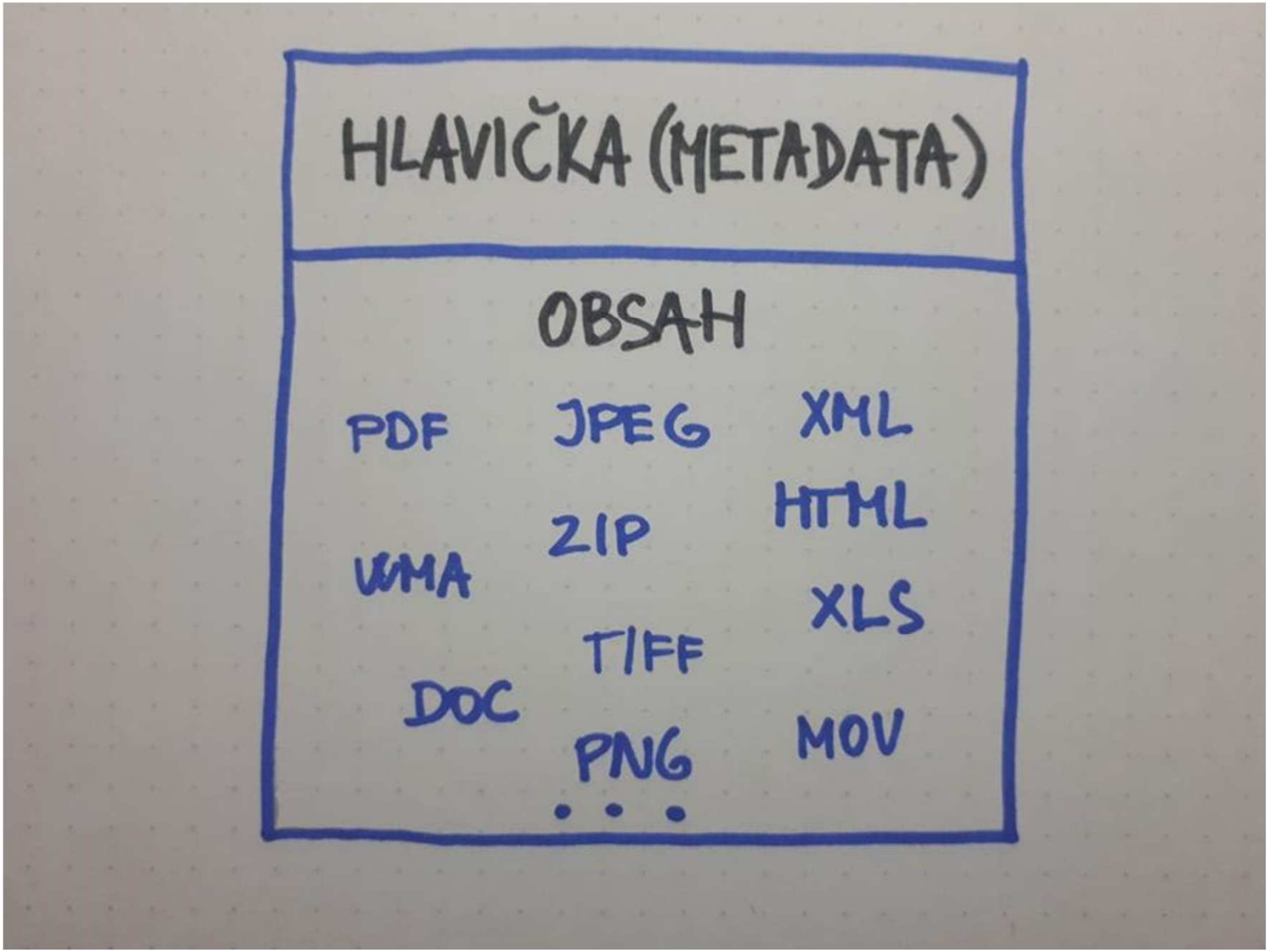
- unikátní klíč
- databázová konstrukce
- slouží ke zrychlení vyhledávání v databázi
- funguje podobně jako rejstřík v knize

| INDEX | PAGE TITLE | URL | TIMESTAMP |
|---|---|---|---|
| 1 | Povodně 2013 \| Webarchiv.cz | https://www.webarchiv.cz/cs/tematicke-kolekce/p | 20. 10. 2019 19:33:23 |
| 2 | Krajské a senátní volby 2016 \| Webarchi | https://www.webarchiv.cz/cs/tematicke-kolekce/k | 20. 10. 2019 19:33:08 |
| 3 | Tematické sbírky \| Webarchiv.cz | https://www.webarchiv.cz/cs/tematicke-kolekce | 20. 10. 2019 19:33:03 |
| 4 | Klíčové slovo masmédia a politika \| Web | https://www.webarchiv.cz/cs/klicove-slovo/3515- | 20. 10. 2019 19:32:57 |
| 5 | ANO, bude líp | https://www.anobudelip.cz/cs/ | 20. 10. 2019 19:32:41 |
| 6 | Vládní krize 2017 \| Webarchiv.cz | https://www.webarchiv.cz/cs/tematicke-kolekce/v | 20. 10. 2019 19:32:35 |
| 7 | Tematické sbírky \| Webarchiv.cz | https://www.webarchiv.cz/cs/tematicke-kolekce | 20. 10. 2019 19:32:31 |
| 8 | Webarchiv podle oborů \| Webarchiv.cz | https://www.webarchiv.cz/cs/katalog-stranek/24- | 20. 10. 2019 19:32:28 |
| 9 | Webarchiv podle oborů \| Webarchiv.cz | https://www.webarchiv.cz/cs/katalog-stranek/25- | 20. 10. 2019 19:32:24 |
| 10 | Webarchiv podle oborů \| Webarchiv.cz | https://www.webarchiv.cz/cs/katalog-stranek/1-a | 20. 10. 2019 19:32:20 |
| 11 | Webarchiv podle oborů \| Webarchiv.cz | https://www.webarchiv.cz/cs/katalog-stranek | 20. 10. 2019 19:32:17 |
| 12 | Památník českého internetu \| Webarchiv | https://www.webarchiv.cz/cs/ | 20. 10. 2019 19:32:15 |
| 13 | O Webarchivu \| Webarchiv.cz | https://www.webarchiv.cz/cs/o-webarchivu | 20. 10. 2019 19:32:11 |
| 14 | Památník českého internetu \| Webarchiv | https://www.webarchiv.cz/cs/ | 20. 10. 2019 19:32:06 |

# *WARC*

- WARC = Web ARChive
- Mezinárodní norma ISO 28500:2017
- struktura záznamu dat
- myšlenkový kontejner, který obsahuje data
- široké spektrum formátů (txt, avi, jpeg, docs, gif, mp3, png, csc, html,…)
- struktura = hlavička + obsah

# *obrázek WARC*



*datový objekt*

- virtuální jednotka, struktura = hlavička + obsah

*sklizeň*

- návod pro crawler, entita nadřazená WARC souborům

*Wayback Machine*

-nástroj pro zobrazení archivovaných stránek

# Webarchiv

http://strazprirody.cz Vyhledat

http://strazprirody.cz byl archivován **64x.** Přejít na první archivovanou verzi: 13. října 2010.

Požadovaný archivovaný objekt může odkazovat na identickou, ale dříve archivovanou verzi. Více na Webarchiv.cz


## 2019

### I

| | | | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 13 | 14 | 15 | 16 | 17 | 18 | 19 |
| 20 | 21 | 22 | 23 | 24 | 25 | 26 |
| 27 | 28 | 29 | 30 | 31 | | |

### II

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | 1 | 2 |
| 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| 17 | 18 | 19 | 20 | 21 | 22 | 23 |
| 24 | 25 | 26 | 27 | 28 | | |

### III

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | 1 | 2 |
| 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| 17 | 18 | 19 | **20** | 21 | 22 | 23 |
| 24 | 25 | 26 | 27 | 28 | 29 | 30 |
| 31 | | | | | | |

### IV

| | | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| 21 | 22 | 23 | **24** | 25 | 26 | 27 |
| 28 | 29 | 30 | | | | |

### V

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 4 |
| 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| 19 | 20 | 21 | 22 | 23 | 24 | 25 |
| 26 | 27 | 28 | 29 | 30 | 31 | |

### VI

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | | 1 |
| 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| 16 | 17 | 18 | 19 | 20 | 21 | 22 |
| 23 | 24 | 25 | 26 | 27 | 28 | 29 |
| 30 | | | | | | |

### VII

| | | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| 21 | 22 | 23 | 24 | 25 | 26 | 27 |
| 28 | 29 | 30 | 31 | | | |

### VIII

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | 1 | 2 | 3 |
| 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 11 | 12 | 13 | 14 | 15 | 16 | 17 |
| 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| 25 | 26 | 27 | 28 | 29 | 30 | 31 |

# *rekapitulace*

- sklízení webových stránek
- tvorba metadat
- uložení obsahu a metadat do WARC souboru
- zobrazení archivovaného obsahu pomocí speciálního softwaru

# *praktická část*

# *instalace Webrecorder player*

- instalace Webrecorder aplikace
- **https://github.com/webrecorder/webrecorder-player/releases/tag/v1.7.0**

# *3 přístupy k obsahu*

- interaktivní - Webrecorder Player
- textový editor
- příkazový řádek

# *Webrecorder*

- open-source nástroj, zdarma
- statický i dynamický obsah
- projekt Rhizome, pod záštitou The Andrew W. Mellon Foundation

**https://webrecorder.io**

**(návod: https://guide.webrecorder.io)**

# *interaktivní prohlížení obsahu*

- v programu WebArchiver Player
- nebo jen v prohlížeči
- fungují jen odkazy, které reálně sklidíme (ostatní nebudou fungovat)
- sklidí se i hlášení o chybě

# *čtení pomocí textového editoru*

- WARC soubor je nutné nejprve odzipovat
  - mac: program The Unarchiver
  - windows: zobrazit v průzkumníku, použít 7-Zip
- poté otevřít v jakémkoli textové editoru

# *čtení pomocí příkazového řádku*

tahák co kam napsat

obě varianty jedna pro mac a jedna pro windows

# *psí kusy*

ukázky práce co jde všechno vytahat z warc souborů a jak...

*a to je vše přátelé*

gratuluju teď jste udělali první krok k datové analýze jej!