Metodika pro tvorbu, uložení a zpřístupnění technických a administrativních metadat z webového archivu

Kvasnica, Jaroslav,; Vozár, Zdenko; Haškovcová, Marie,; Kodad Holoubková, Monika
2020

Národní knihovna České republiky

Webarchiv

Metodika pro tvorbu, uložení a zpřístupnění technických a administrativních metadat z webového archivu

Autoři: Jaroslav Kvasnica, Zdenko Vozár, Marie Haškovcová, Monika Holoubková

Datum: 2020
Verze: Verze 1.0

Metodika vznikla na základě institucionální podpory dlouhodobého koncepčního rozvoje výzkumné organizace poskytované Ministerstvem kultury.

Obsah

OPONENTI

Ing. Martin Lhoták, Knihovna AV ČR, v. v. i.

PhDr. Zdeněk Vašek, Ph.D., Ústav dějin a Archiv UK

I ÚVOD

1. Účel metodiky

Metodika předkládá postup (pravidla) pro strukturu metadatového záznamu a popis technických a administrativních metadat z webového archivu. Nastiňuje teoretický rámec problematiky tvorby metadat webových archivů, popisuje granularitu záznamu a jednotlivá metadatová pole s hodnotami, kterých mohou nabývat, a jejich možnou formou zápisu. Vychází především z metadatové specifikace pro webové archivy mezinárodního konsorcia International Internet Preservation Consortium. Metodika byla vyvinuta pro potřeby Českého webového archivu NK ČR (Webarchiv). Je navržena tak, aby mohla sloužit dalším archivům webového obsahu. Jejím účelem je popis technických a administrativních metadat, která se vztahují k jednotlivým sklizním včetně kontejnerového formátu WARC. Umožňuje tak lépe pracovat s technickými a administrativními metadaty z webových archivů, metadata tak mohou být zapojena do dalších výzkumů. Rozlišeny jsou tři specifikace metadat - první náleží k popisu sklizně, druhá ke kontejnerovému formátu a třetí k indexu (viz kapitola III Metadatový záznam). V souvislosti s touto metodikou byl vyvinut nástroj - software Grainery, který lze použít pro generování, extrakci, evidenci a zobrazení metadat podle postupů stanovených v metodice (viz kapitola 8. Grainery).

2. Určení

Metodika byla vyvinuta na míru potřebám Oddělení archivace webu NK ČR, může sloužit všem webovým archivům a paměťovým institucím, které archivují webový obsah. Jejími uživateli mohou být techničtí správci a kurátoři webových archivů, producenti webového obsahu i běžní uživatelé webových archivů - badatelé i široká veřejnost. Je zaměřena na popis, uložení a standardizaci postupů průzkumu fondu webového archivu, který kombinuje proprietární vlastnosti definic formátu CDX a WARC a filesystémové názvosloví hierarchického úložiště vycházejícího z příkladů ukládání pro WARC dle přílohy C normy ISO (28500:2009(E): WARC file size and name recommendations). V tomto směru je metodika

vhodná pro každého, kdo chce obohatit svůj webový archiv o administrativní a technická metadata.

3. Proprietární jmenné konvence

Část metodiky reaguje na partikulární historii a organizaci hierarchického úložiště NK ČR s dlouhou tradicí (první archivní kopie pocházejí z roku 2001). To je obzvlášť patrné ve zkoumání elementů jako parentDir odrážející prefix sklizně. Také jmenná konvence prefixu sklizně, kterou jsou označovány WARC soubory a na níž tato metodika reaguje, sahá až do počátku českého webového archivu (roky 2003-05). Dle výše citované normy ISO používáme jmennou konvenci pro komprimované WARC soubor, a to: Prefix-Timestamp-Serial-Crawlhost.warc.gz. Ovšem Prefix samotný obsahuje jak typ sklizně, tak i volitelně její pořadí nebo datum a také typy balíků semínek (podle typu sklizní), které byly začleněné do sklizně a často korespondují s její frekvencí. Abstrakce segmentu harvestPrefix zohledňuje zmíněné partikularity, které pro využití v jiných archivech nejsou zdaleka nutné, avšak v metodice zůstávají jako doporučené.

4. Související normy a předpisy

Specifikace vychází ze standardizovaného mezinárodně používaného formátu WARC (https://www.iso.org/standard/44717.html), který užívá k účelům extrakce vyšších metadatových jednotek, tzn. pohybuje se na vyšší sémantické úrovni než uvedený standard.

5. Terminologie

Webová archivace

Archivace webu je proces, který zahrnuje získávání webových zdrojů, jejich ukládání, trvalé uchování, ochranu i jejich zpřístupnění.

Sklízení webu

Proces sběru dat z webu spočívá v automatizovaném mapování, vyhledávání a stahování určitých webových stránek pomocí crawlerů na základě definovaných parametrů. Crawler je speciální počítačový program, který dokáže automaticky procházet a stahovat webové stránky. Používají je nejen internetové vyhledávače, ale i jednotlivé webové archivy.

WARC

Specializovaný kontejnerový formát určený k uložení webových sklizní vytvořených v rámci archivace. Umožňuje agregaci jednotlivých fragmentů staženého webového obsahu, viz https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.0. Sklízení není logicky rozděleno na jednotlivé www servery, probíhá z několika míst zároveň, a ty se pak ukládají do kontejneru, dokud není naplněna jeho kapacita. Po naplnění se kontejner uzavře a začne se plnit nový. Jeden kontejner tedy může obsahovat různé fragmenty z různých www serverů. Jeden WARC může obsahovat řádově tisícovky souborů v nejrůznějších formátech.

Hash

Otisk objektu, kontrolní součet, unikátní řetězec znaků, slouží pro budoucí ověření autenticity nebo poškození obsahu. Metodika používá jako kontrolní hash md5, do budoucna lze uvažovat o dalších metodách výpočtu kontrolních součtů.

Log soubory (logy)

Log soubory (logy) jsou textové soubory obsahující záznamy o činnosti. Generuje je crawler a vztahují se k celé sklizni. Při zpětné analýze slouží k rozpoznání, zda došlo k nějaké chybě, případně k jaké chybě došlo a proč.

Index

Index je databázová konstrukce sloužící ke zrychlení vyhledávacích a dotazovacích procesů v databázi. Index ve formátu CDX je nedílnou součástí webových archivů, představuje kompletní seznam archivních objektů v kontejneru a poté jejich umístění v celém archivu. Používá se k vyhledávání a zobrazování konkrétních záznamů požadovaných koncovým uživatelem. Metadatový záznam zahrnuje informace o indexaci dat.

Metadata

Metadata jsou strukturovaná data, která nesou informace o archivovaných datech. Jejich funkce je popisná, selekční a archivační.

URI

URI (Uniform Resource Identifier) je textový řetězec s definovanou strukturou, který slouží k přesné specifikaci zdroje informací.

II TEORETICKÁ VÝCHODISKA

6. Archivace webu a Webarchiv

Prostředí internetu je rozsáhlou platformou veřejné a soukromé komunikace i sociální interakce. Elektronické dokumenty se staly nositeli kulturního dědictví, které se webové archivy snaží zachovat a ochránit pro další generace. Vzhledem k velké proměnlivosti rychle rostoucího objemu dat (data jsou měněna, přesouvána nebo vymazána) roste i význam archivů webového obsahu, které se stávají cenným pramenem pro studium současnosti a blízké minulosti. Jejich úkolem není jen sběr a dlouhodobé uchovávání dat, ale také jejich zpřístupnění a další využití badatelům i veřejnosti. Přístup k dokumentům však podstatně ovlivňují nejen limity technické, ale i právní či metodické. Vzhledem k legislativním omezením se archivy snaží zveřejňovat alespoň tu část archivu, která není autorsky chráněna, zejména různá metadata, k nimž patří i metadata technická a administrativní, na něž se tato metodika zaměřuje. Hledáním způsobů práce s metadaty se zabývá v mezinárodním měřítku konsorcium webových archivů International Internet Preservation Consortium; hledáním způsobů zpřístupnění metadat na základě analýzy potřeb uživatelů se zabývá také například v rámci organizace OCLC - Online Computer Library Center zvláštní pracovní skupina. Veškerý webový obsah archivovat nelze a vzhledem k rozsáhlosti webu je tak pro jeho archivaci klíčová otázka výběru. Zásadní jsou proto vstupní parametry při archivaci webových stránek, jejich nastavení má dopad na výsledná data v rovině obsahové, formátové i technické. K faktorům, které ovlivňují výslednou podobu archivních dat, patří technická nastavení, strategie výběru zdrojů, tzv. Collection policy, a legislativa. Důležitá jsou i pravidla nakládání s archivními daty - zejména pro jejich mazání a omezování přístupu k obsahu. Při posuzování autenticity archivních kopií webových stránek je nutné zohlednit způsob, jakým byly dokumenty spravovány v okamžiku jejich vytvoření.

Webarchiv, český webový archiv Národní knihovny, plní funkci digitálního archivu webových stránek, soustředí se zejména na archivaci národního webu. Sklízí dokumenty, které jsou na internetu volně dostupné. Zaměřuje se na unikátní elektronické zdroje s dlouhodobou kulturní, vědeckou či historickou hodnotou, které mají bohemikální charakter, tzn. které se k území České republiky vztahují teritoriálně, autorsky, jazykově nebo obsahově. Ve snaze postihnout v co největší míře český web provádí jednou až dvakrát ročně tzv. celoplošnou sklizeň. Jedná se o automatický sběr dat, kdy díky spolupráci se správcem české domény CZ.NIC archivuje všechny webové stránky s národní doménou .cz. Kromě toho provádí sklizně výběrové na základě selekčních kritérií, jejichž cílem je snaha o zachycení vzorku českého kulturního

dědictví na webu napříč všemi oblastmi lidského vědění (používá metodu Konspektu, zdroje jsou sklízeny pravidelně dle kurátory stanovené frekvence), a připravuje kolekce tematické, kdy se snaží zachytit důležitá aktuální celospolečenská témata, která rezonují v prostoru českého webu. Vybrané zdroje jsou registrovány v České národní bibliografii podle katalogizačních pravidel RDA (Resource Description and Access) a struktura zápisu odpovídá mezinárodnímu metadatovému standardu pro popis dokumentů v knihovních informačních systémech MARC (Machine-Readable Cataloging). Jsou dostupné na webových stránkách Webarchivu a v katalogu Národní knihovny. Podrobné informace o strategii budování českého webového archivu viz dokument Collection Policy (https://www.webarchiv.cz/static/www/download/collection-policy.pdf).

7. Metadata

Metadatový popis je klíčovou součástí dat umožňující s daty dále pracovat, vytvářet různé analýzy nebo například definovat množinu dat pro vědecký výzkum. Pokud mají vědci dostatek metadat, může být zdroj, se kterým pracují, považován za relevantní, a zvyšuje tak reprezentativní hodnotu výzkumu. Pro správce, kurátory i uživatele - ať už badatele nebo širokou veřejnost - představují metadata možnost prvotní analýzy archivu, aniž by museli pracovat s obrovským množstvím dat. Administrativní a technická metadata obsahují podrobné informace o souborovém formátu, ve kterém jsou data uložena a popisují technické údaje zaznamenané během sklizní. Vztahují se k sklizním, ke kontejnerovým formátům, v nichž jsou archivovaná data uložena, a k indexu. K základním údajům patří například datum zahájení a ukončení sklizně, její typ, rozsah nebo autor (podrobně viz kapitola III. Metadatový záznam). Princip uložení dat a metadat v kontejnerovém formátu spočívá v tom, že každému datovému objektu předchází hlavička s metadaty. Datovým objektem mohou být buď samotné soubory stažené z webových stránek nebo metadatové záznamy. Metadata, která vznikla pro potřeby webových archivů, umožňují mimo jiné zpětně identifikovat případné chyby - kdy a proč se staly, proč není stránka v archivu správně uložená a podobně. Opatřování dat kvalitními metadaty umožňuje zjistit, co se ve sbírkách nachází, proto jsou pro správu webového archivu a jeho dalšího využití důležitá.

8. Grainery

Grainery je nejen název standardu, ale také software umožňující vygenerování záznamu z archivních dat a následně zobrazení metadat. Cílem je preciznější monitoring úložiště, jeho správa, podklad pro bitovou a logickou ochranu a grafické zpracování dat, které názorně přibližuje jeho obsah správcům, kurátorům a uživatelům. Umožňuje pracovat s jeho obsahem na základě různých analýz metadat, například o původu, vzniku a objemu archivních dat. Aplikace přináší možnost přehledně extrahovat metadata, která se nacházejí na více místech (u archivních dat, u indexačních souborů, část vytváří sklízeč jako soubory s logy). Zejména metadata, která jsou uložena u archivních dat, byla dosud dostupná pouze technickým správcům. Grainery dále přináší možnost tato metadata přehledně zpřístupnit v grafické podobě, případně přes API (rozhraní pro programování aplikací), které je vhodné pro jejich dávkové zpracování. Prezentační vrstva slouží k procházení a vizualizaci metadat uživateli. Aplikace umožňuje sledování technického stavu webových archivů a rychlý náhled a sumarizaci technických parametrů archivních dat.

Aplikace Grainery, pomocí které lze metadata extrahovat a zobrazit, byla speciálně vyvinutá pro extrakci, formátování, ukládání, zobrazování a API dotazování dat standardu. Umožňuje pracovat s technickými a administrativními metadaty z webových archivů, staví na metadatové specifikaci Grainery 0.35 pro webové archivy vycházející z širších specifikací IIPC. Skládá se ze dvou hlavních částí: python extraktor metadat Extarc a prezentační vrstva (Grainery frontend). Komunikují společně přes NoSQL bázi (MongoDB) přes formát JSON, který je definován technickou metadatovou specifikací Grainery. Vývojářská dokumentace softwaru Grainery je k dispozici zde: https://github.com/WebarchivCZ/grainery, https://github.com/WebarchivCZ/grainery/wiki.

III METADATOVÝ ZÁZNAM

9. Formát metadatového záznamu

Metadata jsou uložena ve formátu JSON (JavaScript Object Notation), což je formát navržený pro výměnu dat webovými aplikacemi. Jedná se o způsob zápisu určený pro přenos dat, která mohou být organizována v polích (indexovaných i neindexovaných), nebo agregována v objektech (pole dvojic název/hodnota). Vstupem je datová struktura - číslo, řetězec, boolean (datový typ reprezentovaný jednou ze dvou hodnot: true - pravda nebo false - nepravda), objekt

nebo pole, které je z nich složené. Výstupem je řetězec. Formát může vytvářet člověk, zároveň lze generovat a analyzovat strojově. Je založený na skriptovacím jazyku JavaScript. Datové struktury jsou realizovány v těchto konstrukcích:

Objekt je uvozený levou složenou závorkou a zakončený pravou složenou závorkou. Za každým názvem následuje dvojtečka, dvojice název: hodnota jsou odděleny čárkou.

```
{
    "logs": true,
    "path": "logs/crawl",
    "fileName": ["crawler00.tar.gz", "crawler01.tar.gz", "crawler03.tar.gz"]
      }
```

Pole začíná levou hranatou závorkou a končí pravou hranatou závorkou. Hodnoty jsou odděleny čárkou.

```
["crawler00.tar.gz", "crawler01.tar.gz", "crawler03.tar.gz"]
```

Hodnotou se rozumí řetězec uzavřený do dvojitých uvozovek - číslo, boolean, objekt nebo pole. Tyto struktury mohou být vnořovány.

```
"crawler00.tar.gz"
```

Řetězec tvoří nula nebo více znaků uzavřených do dvojitých uvozovek a využívající únikových sekvencí (escape sequence) s použitím zpětného lomítka.

```
"Narodni knihovna CR"
```

Mezi jednotlivé znaky a hodnoty lze vkládat bílé znaky (whitespace).

10. Struktura metadatového záznamu

Metadatový záznam odpovídá jedné digitální (intelektuální) entitě ve webovém archivu. Těmito entitami jsou sklizeň, archivní kontejner a soubor s indexem (cdx).

Rozlišujeme tedy tři typy metadatových záznamů objektů:

- **Sklizeň (*harvest)***
    - Sklizeň jako zastřešující intelektuální entita je tvořena z kontejnerů. Musí být intelektuální entitou z toho důvodu, že obsah jedné webové stránky je uložen do

více kontejnerů včetně propojeného kontextu, protože žádná webová stránka není osamocená, ale je pomocí hypertextových odkazů součástí větší sítě. Prostřednictvím metadat sklizně lze zjistit vazby mezi kontejnery, potřebné pro zobrazení archivních kopií webových stránek.

- **Kontejner (*container)***
  - Archivní kontejner je balíček dat ve specializovaném archivním formátu WARC. Kontejner obsahuje fragmenty webových stránek, tzn. soubory stažené z webové stránky zabalené do kontejnerového formátu.
- **CDX index (*cdx*)**
  - Databázová konstrukce, která umožňuje zrychlení vyhledávacích a dotazovacích procesů v databázi. V případě webových archivů je index ve standardizovaném formátu CDX, který obsahuje seznam všech URI obohacených o základní metadata.

Každý metadatový záznam se skládá z hlavičky Grainery, která mimo jiné obsahuje definici typu záznamu, který uvozuje. Dále následuje jedna až n sekcí patřící k odpovídajícímu typu záznamu. Celý záznam je zakončen patičkou, která je pro všechny typy stejná, obdobně jako hlavička. Patička pak obsahuje kontrolní údaje o validaci objektu. Popis jednotlivých polí obsahuje název, definici, příklad a určení povinnosti. Vlastnosti, které jsou považovány za signifikantní, jsou označeny jako povinné.

10.1 Hlavička Grainery

Hlavička uvozuje každý metadatový záznam generovaný pomocí Grainery. Obsahuje základní informace o vytvoření metadatového záznamu. Pokud je využit generátor metadat Grainery-Extarc, pak je hlavička vytvořena automaticky. Hlavička je povinnou součásti pro využití zobrazovacího nástroje Grainery a je doporučena pro další interoperabilitu metadatového záznamu.

Pole: recType	
Definice: Typ metadatového záznamu (může nabývat hodnot buď harvest, container nebo cdx).	
Příklad: *harvest*	Povinnost: povinný

Pole: author	
Definice: Tvůrce metadatového záznamu (operátor extrahujícího skriptu).	
Příklad: *NK CR*	Povinnost: povinný

Pole: date	
Definice: Datum vytvoření metadatového záznamu ve tvaru ISO 860. Ve formátu rok/měsíc/den hodiny/minuty/vteřiny [YYYY-MM-DDTHH:MM:SSZ].	
Příklad: *2019-11-14T19:10:46.983Z*	Povinnost: povinný

Pole: standard	
Definice: Název verze použitého metadatového záznamu. Aktuální verze standardu je vždy publikována v repozitáři https://github.com/WebarchivCZ/grainery. Verze popsaná v metodice je verzí 1.0	
Příklad: *Grainery 0.4*	Povinnost: povinný

Příklad hlavičky Grainery ve formátu JSON


```
{
    "_id": "5dcd98a695bcf5a1a194f0be",
    "recType": "harvest",
    "author": "NKCR",
    "date": "2019-11-14T19:10:46.983Z",
    "standard": "Grainery 0.4",
}
```


10.2 Revision (Patička Grainery)

Patičku tvoří informace o kontrolním součtu kontejneru md5. Je ukazatelem, zda je kontejner v pořádku, jestli nebyl zničen nebo poškozen, kdy a jaká validace ho čeká. Každý metadatový záznam končí touto patičkou. Patička je uvozena v sekci *revision*.

Pole: dateOfValidation	
Definice: Datum, kdy proběhla poslední kontrola integrity dat a jaký byl výsledek (ve formátu rok/měsíc/den hodiny/minuty/vteřiny [YYYY-MM-DDTHH:MM:SSZ]).	
Příklad: *2019-12-04T19:10:46.983Z*	Povinnost: povinný

Pole: statusOfValidation	
Definice: Výsledek kontroly integrity dat. Může nabývat hodnoty FIRST / FIRST-FAILED / VALIDATED / TOBEVALIDATED / FAILED. FIRST znamená, že validace ještě nebyla uskutečněná, ale soubor byl zachycen v rámci extrakce z úložiště. FIRST-FAILED popisuje, že soubor byl zachycen, přečten, ale nezdařila se tvorba jeho hashe (otisku). VALIDATED znamená, že soubor byl validován a je v garanční době, momentálně nastavené na 700 dní. TOBEVALIDATED znamená, že soubor byl validován, ale zbývá posledních třicet dní do vypršení garanční doby. FAILED soubor byl validován oproti hashOrig, ale hashLast je odlišný, případně se md5 nezdařil vůbec (např z důvodu nedostupného úložiště) a pole hashLast má hodnotu NA = not accessible.	
Příklad: *VALIDATED*	Povinnost: povinný

Pole: nextLastDateOfValidation	
Definice: Datum a čas, do kdy bude provedena další kontrola. Standardně +730 dní. V závislosti na něm se nastavuje statusOfValidation (ve formátu rok/měsíc/den hodiny/minuty/vteřiny [YYYY-MM-DDTHH:MM:SSZ]).	
Příklad: *2021-12-03T19:10:46.983Z*	Povinnost: povinný

Pole: hashOrig	
Definice: Hash (otisk) objektu, kontrolní součet, unikátní řetězec znaků, slouží pro budoucí ověření autenticity nebo poškození obsahu. Označuje hash souboru při vytvoření záznamu.	
Příklad: *b3cd915d758008bd19d0f2428fbb354a '}*	Povinnost: povinný

Pole: hashLast	
Definice: Hash (otisk) objektu, kontrolní součet, unikátní řetězec znaků, slouží pro budoucí ověření autenticity nebo poškození obsahu. Označuje poslední verzi ověřovacího součtu.	
Příklad: *2db95e8e1a9267b7a1188556b2013b33*	Povinnost: povinný

Commentaries

Pole: exist	
Definice: Informace o tom, jestli komentář existuje. Toto pole je určeno pro komentáře ke konkrétním záznamům v bázi, informace o úpravách. Může nabývat hodnot typu boolean: true - komentář byl vložen false - komentář nebyl vložen	
Příklad: *true*	Povinnost: povinný

Pole: text	
Definice: Komentáře k záznamu, informace o úpravách. Například NA = not accessible.	
Příklad: *NA*	Povinnost: povinný

Příklad patičky Grainery ve formátu JSON


```
},
        "revision": {
          "dateOfValidation": "2019-12-04T19:10:46.983Z",
          "statusOfValidation": FAILED,
          "nextLastDateOfValidation": "2021-12-03T19:10:46.983Z",
          "hashOrig": "b3cd915d758008bd19d0f2428fbb354a",
          "hashLast": "2db95e8e1a9267b7a1188556b2013b33",
          "commentaries": { "exists": false, "text": "NA" }
        }
```


10.3 Harvest - sklizeň (recType: harvest)

Harvest (sklizeň) je proces sběru dat z webu, který spočívá v automatizovaném mapování, vyhledávání a stahování určitých webových stránek pomocí crawlerů (speciálních počítačových programů) na základě definovaných parametrů. Typ záznamu harvest představuje abstrakci z metadatových toků ostatních archivovaných souborů. Jde o vyšší jednotící vrstvu uchovávající důležité záznamy pro plnou rekonstrukci sklizně a jejích parametrů.

Harvest se skládá ze sekcí Hlavička Grainery, Harvest, HarvestCrawl, Paths, Revision (Patička Grainery):

10.3.1 Hlavička Grainery

(viz Hlavička Grainery 10.1)

10.3.2 Harvest

HarvestPrefix

Pole: harvestNameStand	
Definice: Standardizované unikátní jméno sklizně (abstrakce viz níže). Při odkazování na sklizeň má však jako klíč přednost pole harvestID. Název sklizně je definovaný WARC soubory. Soubory definující stejnou sklizeň se řadí jako její součásti.	
Příklad: *V6M_2017-10-V6M_2017-10-05*	Povinnost: povinný

Pole: harvestFromWarcinfo	
Definice: Agregovaný záznam názvů, list o jednom nebo více elementech přejatý ze sekce container - isPartOf dle ISO 28500 pro warcinfo (viz container - isPartOf). Od položek harvestNameFNtrunc a harvestDirsName se může lišit, nebo obsahovat více záznamů. Je prioritní pro další účely zpracování. Pokud by ho kolekce neobsahovala, abstrahuje se pro vyplnění záznamu Harvest dle vnějšího příznaku harvestNameFNtrunc. V případě nekonzistence je nutný manuální průzkum úložiště.	
Příklad: *V6M_2017-10-V6M_2017-10-05*	Povinnost: doporučený

Pole: harvestNameFNtrunc	
Definice: Agregovaný záznam názvů, list o 1 nebo více elementech přejatý z WARC- FileName (viz container - FileName).	
Příklad: *V6M_2017-10-V6M_2017-10-05*	Povinnost: doporučený

Pole: harvestDirsName	
Definice: Agregovaný záznam názvů, list o jednom nebo více elementech názvu přejatý z bloku paths, element basic - parentDir (viz container - path).	
Příklad: *V6M_2017-10-05*	Povinnost: doporučený

Pole: harvestType	
Definice: Označuje, o jaký typ sklizně se jedná, například o měsíční sklizeň, celoplošnou sklizeň, jednorázovou sklizeň, testovací apod. Povolená pole: Topics - speciální tematická sklizeň Serials - pravidelná sklizeň (kombinace výběrových sklizní s různou frekvencí sklízení) Continuous - pravidelná sklizeň s denní nebo nižší frekvencí Totals - celoplošná sklizeň domény .cz Tests - testovací sklizeň Requests - sklizeň vyžádaná jinou institucí	
Příklad: *Serials*	Povinnost: doporučený

Pole: harvestSuffix[harvestFreq]	
Definice: Frekvence a typ sklizně může obsahovat jedno a více polí (list). Pole může nabývat kombinace hodnot frekvence a typu, označený jako harvestFreq: V - výběrová sklizeň T - tematická sklizeň 1M - měsíční frekvence 2M - dvouměsíční frekvence 3M - čtvrtletní frekvence 6M - šestiměsíční frekvence 12M - roční frekvence V-1 - jednorázová sklizeň CZ18 - celoplošná sklizeň ArchiveIt - jednorázová sklizeň nových semínek OneShot - další přidaná jednorázová semínka ve formě: datum RRRR-MM(-DD) + název - typicky zejména pro typ sklizně Topics, Tests a Requests, protože se nesklízí pravidelně podle typu frekvence, důležitá je časová specifikace	
Příklad: *"V6M", "2017-10-05"*	Povinnost: doporučený

Pole: date	
Definice: Datum zahájení sklizně (ve formátu rok/měsíc/den hodiny/minuty/vteřiny [YYYY-MM-DDTHH:MM:SSZ]).	
Příklad: *2014-05-15T15:52:20Z*	Povinnost: doporučený

Pole: harvestID	
Definice: Jedinečné označení (identifikátor) sklizně (URN = Uniform Resource Name) sloužící k jednoznačné identifikaci sklizně, jeho pomocí lze na obsah odkazovat, typ identifikátoru podle normy UUID. Je počítán jako UUID5 dle NAMESPACE_DNS a názvu sklizně pomocí algoritmu SHA1. Je tedy zpětně rozklíčovatelný, což zvyšuje jeho sémantickou hodnotu.	
Příklad: *621f785e-9401-4be2-bb48-1039960748fc*	Povinnost: povinný

Pole: size	
Definice: Velikost sklizně v bytech, součet velikostí všech komprimovaných kontejnerů.	
Příklad: *618029*	Povinnost: doporučený

Pole: warcsNumber	
Definice: Počet WARC souborů.	
Příklad: *12092*	Povinnost: povinný

10.3.3 HarvestCrawl

Pole: logs	
Definice: Log soubory (logy) jsou textové soubory obsahující záznamy o činnosti. Při zpětné analýze slouží k rozpoznání, zda došlo k nějaké chybě, případně k jaké chybě došlo a proč. Může nabývat hodnot typu boolean: true - log byl vytvořen false - log nebyl uchován/dohledán	
Příklad: *true*	Povinnost: povinný

Pole: path	
Definice: Cesta k logům na úložišti.	
Příklad: *logs/crawl*	Povinnost: nepovinný

Pole: fileName	
Definice: Seznam názvů souborů obsahujících logy.	
Příklad: *"crawler00.tar.gz", "crawler01.tar.gz", "crawler03.tar.gz"*	Povinnost: nepovinný

10.3.4 Paths

Pole: cdxsID	
Definice: List jednotlivých WARC souborů patřících recType: harvest. Zdroj cdx - cdxID.	
Příklad: *105f23c9-b037-4c1d-901c-dcf272877d9f*	Povinnost: doporučený

Pole: warcsID	
Definice: List jednotlivých WARC souborů patřících recType: harvest. Zdroj container - warcID.	
Příklad: *105f23c9-b037-4c1d-901c-dcf272877d9f*	Povinnost: povinný

Pole: warcsFilenames	
Definice: List nekrácených Filenames jednotlivých WARC patřících recType: harvest.	
Příklad: *V6M_2017-10-05-crawler00.webarchiv.cz-warc.gz*	Povinnost: povinný

10.3.5 Revision (Patička Grainery)

viz Revision (Patička Grainery) 10.2

10.4 Container (recType: container)

Kontejnerový formát umožňuje agregaci jednotlivých fragmentů staženého webového obsahu, viz https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.0. Konstrukce záznamů typu harvest z něho vychází a je podle všech participujících záznamů typu container ověřována.

Container se skládá ze sekcí Hlavička Grainery, Container, Type, Paths, Revision (Patička Grainery):

10.4.1 Hlavička Grainery

(viz Hlavička Grainery 10.1)

10.4.2 Container

Pole: fileName	
Definice: Název kontejnerového souboru na úložišti.	
Příklad: *V6M_2017-10-05-crawler00.webarchiv.cz-warc.gz*	Povinnost: povinný

Pole: warcID	
Definice: Jedinečné označení (identifikátor) kontejneru sloužící k jednoznačné identifikaci, jeho pomocí lze na obsah odkazovat, typ identifikátoru podle normy UUID.	
Příklad: *3d1ee065-1ae5-469f-bc02-1a8b3bcbe6b*	Povinnost: povinný

Pole: isPartOf	
Definice: Informace o tom, k jaké sklizni kontejner náleží.	
Příklad: *V6M_2017-10-05*	Povinnost: povinný

Pole: hostName	
Definice: Název stroje, kterým byl kontejner vytvořen.	
Příklad: *crawler00.webarchiv.cz*	Povinnost: doporučený

Pole: ip	
Definice: IP adresa stroje, kterým byl kontejner vytvořen.	
Příklad: *10.3.0.23*	Povinnost: doporučený

Pole: contentLenght	
Definice: Velikost obsahu v bytech.	
Příklad: *955*	Povinnost: povinný

Pole: operator	
Definice: Jméno toho, kdo kontejner vytvořil.	
Příklad: *Jan Novák*	Povinnost: povinný

Pole: publisher	
Definice: Instituce odpovědná za uskutečněný sběr.	
Příklad: *Narodni knihovna CR*	Povinnost: povinný

Pole: audience	
Definice: Pro koho je sklizeň určená, například pro čtenáře Národní knihovny. Jiné publikum je uváděno v případě sklizně pro spolupracující instituci.	
Příklad: *Narodni knihovna CR users*	Povinnost: povinný

Pole: robots	
Definice: Informace o tom, zda je respektována restrikce robots.txt. Může nabývat hodnot: classic, ignore, custom, most-favored, most-favored-set. classic - dodržování pravidel robots.txt pro nakonfigurovaného user-agenta ignore - ignorování pravidel robots.txt custom - dodržování pravidel dle rozhodnutí operátora most-favored - procházení URI pokud robots.txt užitému typu klienta (user-agentovi) umožňuje sklízení most-favored-set - vyžaduje dodání sady user-agentů a zvolení nejvhodnějšího	
Příklad: *ignore*	Povinnost: povinný

Pole: dateOfOrigin	
Definice: Datum vytvoření kontejneru (ve formátu rok/měsíc/den hodiny/minuty/vteřiny [YYYY-MM-DDTHH:MM:SSZ]).	
Příklad: *2017-10-05T22:31:23.000Z*	Povinnost: povinný

Pole: size	
Definice: Velikost WARC souborů v bytech.	
Příklad: *1055*	Povinnost: povinný

10.4.3 Type

Pole: format	
Definice: Formát kontejnerového souboru - WARC.	
Příklad: *WARC File Format 1.0*	Povinnost: doporučený

Pole: conformsTo	
Definice: Standard, podle jakého se formát řídí.	
Příklad: *https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.0/*	Povinnost: povinný

Pole: warcType	
Definice: Způsob popisu záznamu kontejneru.	
Příklad: *warcinfo*	Povinnost: povinný

Pole: mimeTypeXML	
Definice: Označení souborového formátu na internetu.	
Příklad: *application/warc*	Povinnost: povinný

10.4.4 Paths

Basic

Pole: absolute	
Definice: Absolutní cesta k fyzickému uložení na úložišti. Prozatimní údaj.	
Příklad: */mnt/archive/13/serials/Serials-2013-07-1M_ArchiveIt'*	Povinnost: povinný

Pole: parentDir	
Definice: Mateřská složka obsahující daný soubor. Dle názvové konvence je často pojmenovaná názvem sklizně. Pro určení názvu je to údaj ale jen indikativní, díky různým operacím na úložišti a lidskému faktoru při nastavování sklizně.	
Příklad: *V6M_2017-10-05*	Povinnost: povinný

Pole: harvestID	
Definice: Jedinečné označení sklizně, které slouží k její jednoznačné identifikaci.	
Příklad: *105f23c9-b037-4c1d-901c-dcf272877d9f*	Povinnost: doporučený

Pole: cdxID	
Definice: Jedinečné označení indexu, které slouží k jednoznačné identifikaci.	
Příklad: *105f23c9-b037-4c1d-901c-dcf272874d9f*	Povinnost: doporučený

10.4.5 Revision (Patička Grainery)

viz Revision (Patička Grainery) 10.2

10.5 CDX (recType: cdx)

Index je databázová konstrukce, sloužící ke zrychlení vyhledávacích a dotazovacích procesů v databázi. Index ve formátu CDX je nedílnou součástí webových archivů, představuje kompletní seznam archivních objektů v kontejneru a poté jejich umístění v celém archivu. Metadatový záznam zahrnuje informace o indexaci dat.

Index se skládá ze sekcí Hlavička Grainery, CDX, Revision (Patička Grainery):

10.5.1 Hlavička Grainery

(viz Hlavička Grainery 10.1)

10.5.2 CDX

Pole: fileName	
Definice: Název indexového souboru na úložišti.	
Příklad: *V6M_2017-10-05-crawler00.webarchiv.cz-warc.gz.cdx*	Povinnost: povinný

Pole: warcName	
Definice: Název kontejnerových souborů WARC.	
Příklad: *V6M_2017-10-05-crawler00.webarchiv.cz-warc.gz*	Povinnost: doporučený

Pole: exists	
Definice: Informace o tom, jestli index existuje. true - existuje false - neexistuje	
Příklad: *true*	Povinnost: doporučený

Pole: cdxID	
Definice: Jedinečné označení indexu, které slouží k jednoznačné identifikaci sklizně.	
Příklad: *105f23c9-b037-4c1d-901c-dcf272874d9f*	Povinnost: povinný

Pole: size	
Definice: Velikost indexu v bytech.	
Příklad: *158*	Povinnost: povinný

Pole: columns	
Definice: Počet sloupců indikující rozdílné verze indexu vyvolané použitím různých verzí indexátoru.	
Příklad: *9*	Povinnost: povinný

Pole: lines	
Definice: Počet souborů v kontejneru.	
Příklad: *967062*	Povinnost: povinný

10.5.3 Paths

Basic

Pole: absolute	
Definice: Absolutní cesta k fyzickému uložení na úložišti. Prozatimní údaj.	
Příklad: */mnt/archive/13/serials/Serials-2013-07-1M_ArchiveIt'*	Povinnost: povinný

Pole: parentDir	
Definice: Mateřská složka obsahující daný soubor. Dle názvové konvence je často pojmenovaná názvem sklizně. Pro určení názvu je to údaj ale jen indikativní, díky různým operacím na úložišti a lidskému faktoru při nastavování sklizně.	
Příklad: *index*	Povinnost: povinný

Pole: warcID	
Definice: Jedinečné označení (identifikátor) kontejneru, sloužící k jednoznačné identifikaci, jeho pomocí lze na obsah odkazovat, typ identifikátoru podle normy UUID. Užívá se u recType: cdx, zdroj container - warcID.	
Příklad: *3d1ee065-1ae5-469f-bc02-1a8b3bcbe6b2*	Povinnost: doporučený

Pole: harvestID	
Definice: Jedinečné označení (identifikátor) sklizně sloužící k jednoznačné identifikaci sklizně, jeho pomocí lze na obsah odkazovat, typ identifikátoru podle normy UUID. Užívá se u recType: container, cdx, zdroj harvest - harvestID.	
Příklad: *621f785e-9401-4be2-bb48-1039960748fc*	Povinnost: doporučený

10.5.4 Revision (Patička Grainery)

viz Revision (Patička Grainery) 10.2

10.6 Příklad záznamu

Příklad záznamu ve formátu JSON

10.6.1 Harvest record type


```
{
    "_id": "5dcd98a695bcf5a1a194f0be",
    "recType": "harvest",
    "author": "NKCR",
    "date": "2019-11-14T19:10:46.983Z",
    "standard": "Grainery 0.4",
    "harvest": {
      "harvestPrefix:": {
        "harvestNameStand": "V6M_2017-10-05",
        "harvestFromWarcinfo": "V6M_2017-10-05",
        "harvestNameFNtrunc": "V6M_2017-10-05",
        "harvestDirsName": "V6M_2017-10-05",
        "harvestType": "Serials",
        "harvestSuffix": ["V6M", "2017-10-05"]
      },
      "date": "2017-10-05T11:26:00.000Z",
      "harvestID": "105f23c9-b037-4c1d-901c-dcf272877d9f",
      "size": 618029,
      "warcsNumber": 12092
    },
    "harvestCrawl": {
      "logs": true,
      "path": "logs/crawl",
      "fileName": ["crawler00.tar.gz", "crawler01.tar.gz", "crawler03.tar.gz"]
    },
    "paths": {
      "cdxsID": ["105f23c9-b037-4c1d-901c-dcf272874d9f"],
      "warcsID": ["105f23c9-b037-4c1d-901c-dcf272877d9f"],
      "warcsFilenames": ["V6M_2017-10-05-crawler00.webarchiv.cz-warc.gz"]
    },
    "revision": {
      "dateOfValidation": "2019-12-04T19:10:46.983Z",
      "statusOfValidation": "FAILED",
      "nextLastDateOfValidation": "2021-12-03T19:10:46.983Z",
      "hashOrig": "b3cd915d758008bd19d0f2428fbb354a",
      "hashLast": "2db95e8e1a9267b7a1188556b2013b33",
      "commentaries": { "exists": false, "text": "NA" }
    }
}
```

10.6.2 WARC record type


```
{
  "_id": "5dcd98a695bcf5a1a194f0bf",
  "recType": "container",
  "author": "NKCR",
  "date": "2019-11-14T19:10:46.991Z",
  "standard": "Grainery 0.4",
  "container": {
    "fileName": "V6M_2017-10-05-crawler00.webarchiv.cz-warc.gz",
    "warcID": "9237121d-e513-4352-8591-62637f3ed896",
    "isPartOf": "V6M_2017-10-05",
    "hostName": "crawler00.webarchiv.cz",
    "ip": "10.3.0.23",
    "contentLength": 955,
    "operator": "Zdenko Vozar",
    "publisher": "Narodni knihovna CR",
    "audience": "Narodni knihovna CR users",
    "robots": "ignore",
    "dateOfOrigin": "2017-10-05T22:31:23.000Z",
    "size": 1055
  },
  "type": {
    "format": "WARC File Format 1.0",
    "conformsTo": "https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.0/",
    "warcType": "warcinfo",
    "mimeTypeXML": "application/warc"
  },
  "paths": {
    "basic":{
      "absolute": "/mnt/archive/17/topics/V6M_2017-10-05",
      "parentDir": "V6M_2017-10-05"
    },
    "harvestID": "105f23c9-b037-4c1d-901c-dcf272877d9f",
    "cdxID" : "105f23c9-b037-4c1d-901c-dcf272874d9f"
  },
  "revision": {
    "dateOfValidation": "2019-12-04T19:10:46.991Z",
    "statusOfValidation": FAILED,
    "nextLastDateValidation": "2021-12-03T19:10:46.991Z",
    "hashOrig": "12079a7def9ab255815f54b67147a62d",
    "hastLast": "92eb5ffee6ae2fec3ad71c777531578f",
    "commentaries": { "exists": false, "text": "NA" }
  }
}
```

10.6.3 CDX record type


```
{
    "_id": "5dcd98a695bcf5a1a194f0c0",
    "recType": "cdx",
    "author": "NKCR",
    "date": "2019-11-14T19:10:46.991Z",
    "standard": "Grainery 0.4",
    "cdx": {
      "fileName": "V6M_2017-10-05-crawler00.webarchiv.cz-warc.gz.cdx",
      "warcName": "V6M_2017-10-05-crawler00.webarchiv.cz-warc.gz",
      "exists": true,
      "cdxID" : "105f23c9-b037-4c1d-901c-dcf272874d9f",
      "size": 158,
      "columns": 9,
      "lines": 967062
    },
    "paths": {
      "basic":{
        "absolute": "/mnt/archive/17/topics/V6M_2017-10-05/logs/index",
        "parentDir": "index"
      },
      "warcID": "105f23c9-b037-4c1d-901c-dcf272877d9f",
      "harvestID": "105f23c9-b037-4c1d-901c-dcf272877d9f"
    },
    "revision": {
      "dateOfValidation": "2019-12-04T19:10:46.991Z",
      "statusOfValidation": FAILED,
      "nextLastDateValidation": "2021-12-03T19:10:46.991Z",
      "hashOrig": "4a8a08f09d37b73795649038408b5f33",
      "hastLast": "03c7c0ace395d80182db07ae2c30f034",
      "commentaries": { "exists": false, "text": "NA" }
    }
}
```

IV DOPORUČENÁ LITERATURA A ZDROJE

Archival Science: International Journal on Recorded Information [online]. [cit. 2020-09-21]. ISSN 1573-7500. Dostupné z: https://www.springer.com/journal/10502/

AiOR: Association of Internet Researchers [online]. Association of Internet Researchers, c2020 [cit. 2020-09-22]. Dostupné z: https://aoir.org

BRÜGGER, Niels a Ralph SCHROEDER, ed. *The Web as History*. London: UCL Press, 2017. ISBN 9781911307563. Dostupné také z: https://discovery.ucl.ac.uk/id/eprint/1542998/1/The-Web-as-History.pdf

CUBR, Ladislav. *Autenticita a digitální informace*. Praha, 2017. Dizertační práce. Univerzita Karlova, Filozofická fakulta, Ústav informačních studií a knihovnictví. Vedoucí práce Ivánek, Jiří.

CUBR, Ladislav. *Dlouhodobá ochrana digitálních dokumentů.* Praha: Národní knihovna České republiky, 2010. ISBN 978-80-7050-588-5.

dpc: Digital Preservation Coalition [online]. Digital Preservation Coalition, c2020 [cit. 2020-09-22]. Dostupné z: https://www.dpconline.org/

DN: Documenting the Now [online]. Documenting the Now, c2020 [cit. 2020-09-22]. Dostupné z: https://www.docnow.io/

DOOLEY, Jackie a Kate BOWERS. *Descriptive Metadata for Web Archiving: Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group*. Dublin, Ohio, USA: OCLC Research, 2018. ISBN 978-1-55653-016-6. Dostupné také z: https://www.oclc.org/content/dam/research/publications/2018/oclcresearch-wam-recommendations.pdf

FOLTÝN, Tomáš, Marie HAŠKOVCOVÁ a Andrea PROKOPOVÁ. Vývoj centralizovaného rozhraní pro vytěžování velkých dat z webových archivů.: Základní metodická východiska výzkumného projektu zabývajícího se datovými zdroji českého webu. *ITlib: Informačné technológie a knižnice* [online]. 2020, 20(1), 48 - 51 [cit. 2020-09-21]. Dostupné z: https://itlib.cvtisr.sk/buxus/docs//2020/1_2020/11_1.pdf

HAŠKOVCOVÁ, Marie, Monika HOLOUBKOVÁ, Jaroslav KVASNICA a Markéta HRDLIČKOVÁ. The Acquisition of Czech Web Resources. *Acta Musei Nationalis Pragae – Historia* [online]. 2017, 71(3-4), 41-46 [cit. 2019-11-25]. DOI: 10.2478/amnh-2017-0017. ISSN 2570-6853. Dostupné z: https://content.sciendo.com/view/journals/amnh/71/3-4/article-p41.xml

INTERNATIONAL INTERNET PRESERVATION CONSORTIUM - IIPC [online]. International Internet Preservation Consortium, c2019 [cit. 2019-11-20]. Dostupné z: http://netpreserve.org/

International Internet Preservation Consortium: netpreserveblog [online]. International Internet Preservation Consortium, c2020 [cit. 2020-09-22]. Dostupné z: https://netpreserveblog.wordpress.com/

Internet Histories: Digital Technology, Culture and Society [online]. [cit. 2020-09-21]. ISSN 2470-1483. Dostupné z: https://www.tandfonline.com/toc/rint20/current

ISO 28500:2009. Information and documentation — WARC file format. 1 st ed., 2009. 28 s.

Katalogizační politika: (katalogizace novodobých dokumentů vydaných od roku 1801). *Národní knihovna České Republiky* [online]. Praha: Národní knihovna České Republiky, 03.01.2018 [cit. 2019-11-25]. Dostupné z: https://www.nkp.cz/o-knihovne/odborne-cinnosti/zpracovani-fondu/katalogizacni-politika/

Konspekt: FONDY NÁRODNÍ KNIHOVNY ČR [online]. Praha: Národní knihovna ČR, c2002 [cit. 2019-11-25]. Dostupné z: http://konspekt.nkp.cz/

KVASNICA, Jaroslav a Rudolf KREIBICH. Formátová analýza sklizených dat v rámci projektu WebArchiv NK ČR. *ProInflow: Časopis pro informační vědy* [online]. 2013, **5**(2), 168-177 [cit. 2019-11-20]. ISSN 1804–2406. Dostupné z: https://www.phil.muni.cz/journals/index.php/proinflow/article/view/2013-2-14/911

KVASNICA, Jaroslav, Andrea PROKOPOVÁ, Zdenko VOZÁR a Zuzana KVAŠOVÁ. Analýza českého webového archivu: Provenience, autenticita a technické parametry. *ProInflow: Časopis pro informační vědy* [online]. 2019, **11**(1), 3-21 [cit. 2019-11-20]. ISSN 1804–2406. Dostupné z: http://www.phil.muni.cz/journals/index.php/proinflow/article/view/2019-1-2

KVASNICA, Jaroslav, Barbora RUDIŠINOVÁ a Rudolf KREIBICH. Vědecké využití dat z webových archivů. *Knihovna: knihovnická revue* [online]. 2016, **27**(2), 24-34 [cit. 2019-11-20]. Dostupné z: https://knihovnarevue.nkp.cz/archiv/dokumenty/2016-2/Kvasnica.pdf

KVASNICA, Jaroslav, Barbora RUDIŠINOVÁ, Marie HAŠKOVCOVÁ, Monika HOLOUBKOVÁ a Markéta HRDLIČKOVÁ. Strategie budování sbírky Webarchivu: aktualizované znění. *Webarchiv: památník českého internetu* [online]. Praha: Oddělení archivace webu Národní knihovna ČR, červenec 2019 [cit. 2020-9-20]. Dostupné z: https://www.webarchiv.cz/static/www/download/collection-policy.pdf

KVASNICA, Jaroslav. *Dlouhodobé uchování webového obsahu*. Praha, 2016. Dostupné také z: https://dspace.cuni.cz/bitstream/handle/20.500.11956/82967/DPTX_2012_1_11210_0_345026_0_129352.pdf?sequence=1&isAllowed=y. Diplomová práce. Univerzita Karlova v Praze Filozofická fakulta Ústav informačních studií - studia nových médií. Vedoucí práce PhDr. Mgr. Jan Pokorný, Ph.D.

MASANÈS, Julien, ed. Web Archiving. New York: Springer-Verlag Berlin Heidelberg, 2006. ISBN 978-3-540-23338-1.

TMG: Journal for Media History [online]. Netherlands Institute for Sound and Vision [cit. 2020-09-21]. ISSN 2213-7653. Dostupné z: https://www.tmgonline.nl/

Úvod do JSON. [online] [cit. 2019-11-25]. Dostupné z: https://www.json.org/json-cz.html

Web Archiving Metadata Working Group. *OCLC: Research* [online]. OCLC, c2019 [cit. 2019-11-20]. Dostupné z: https://www.oclc.org/research/themes/research-collections/wam.html

Web Science and Digital Libraries Research Group: Research and Teaching Updates from the Web Science and Digital Libraries Research Group (@WebSciDL) at Old Dominion University. [online]. Web Science and Digital Libraries Research Group at Old Dominion University [cit. 2020-09-23]. Dostupné z: https://ws-dl.blogspot.com/

Web Archiving Section. Society of American Archivists [online]. Society of American Archivists, c2020 [cit. 2020-09-22]. Dostupné z: https://webarchivingrt.wordpress.com/

Webarchiv: památník českého internetu [online]. Praha: Oddělení archivace webu Národní knihovna ČR [cit. 2019-11-25]. Dostupné z: https://www.webarchiv.cz/cs/

WS-DL at ODU-CS. Web Science and Digital Libraries Research Group in the Department of Computer Science at Old Dominion University [online]. Norfolk: Old Dominion University, c2020 [cit. 2020-09-22]. Dostupné z*:* https://ws-dl.cs.odu.edu/

#webarchive. In: *Twitter* [online]. [cit. 2020-09-23]. Dostupné z: https://twitter.com/hashtag/webarchive?src=hashtag_click

#webarchiving. In: *Twitter* [online]. [cit. 2020-09-23]. Dostupné z: https://twitter.com/hashtag/webarchiving?src=hashtag_click