DOI: 10.31915/NWS.2025.21

A MAGYAR WEBTÉR ARATÁSÁVAL KAPCSOLATOS KURÁTORI FELADATOK

Kalcsó Gyula
Magyar Nemzeti Múzeum Közgyűjteményi Központ Országos Széchényi Könyvtár
Digitális Bölcsészeti Központ, Digitális Filológiai és Webarchiválási Osztály
kalcso.gyula@oszk.hu

Abstract

Curatorial Tasks Related to the Harvesting of the Hungarian Web Domain

According to a government decree, the national library's essential task is to carry out a harvest as complete as possible of the Hungarian web domain twice a year and to keep a register of the sites known. This complex task is carried out by the web archiving team of the Digital Philology and Web Archiving Department of the Digital Humanities Centre of the Hungarian National Széchényi Library.

This paper will describe the most important curatorial activities related to this mandated task. It will illustrate the process of registering websites and the methodology for collecting seed URLs. Since the launch of the Hungarian Web Archive in 2017, the number of registered sites has grown significantly. New URLs have been identified from our own harvests, recommendations have been received, and cooperation has been achieved with the Internet Archive being the main source of new URLs.

The seed URL lists need to be maintained before the two annual harvests, which is a complex process involving many steps. The first step is to extract the URLs from the previous captures and sort out those that are not yet known. We automatically retrieve the HTTP status code to determine which sites are live, then retrieve the value of the title tag in the HTML head tag, and see whether the site has a robots.txt file. Based on the structure of the URLs and the information obtained, we can classify the new URLs into the appropriate list. The status codes, the title data as well as the robots.txt are checked for the previously harvested URLs as well, the inactive sites are removed from the lists, and the URLs are classified into the appropriate seed list.

Kulcsszavak: webarchiválás, born digital, a magyar web aratása, webkurátori feladatok

Keywords: web archiving, born digital, harvesting of the Hungarian web, curatorial tasks in web archiving

Mi a magyar webtér?

A magyar webtér szűkebb értelemben az országkód szerinti legfelső szintű doméntartomány, angolul top-domain-level (.hu) tartalma. Tágabb értelemben azonban ide tartozik minden magyar nyelvű webes tartalom, de lehetséges egy még tágabb értelmezés is: a hungarikumnak minősíthető webes tartalmak köre. A muzeális intézményekről, a nyilvános könyvtári ellátásról és a közművelődésről szóló 1997. évi CXL. törvény értelmében „hungarikum: a Magyarország mindenkori területén megjelent minden, továbbá a külföldön magyar nyelven, magyar szerzőtől, illetőleg magyar vonatkozású tartalommal keletkezett valamennyi dokumentum" (Kult tv. 1. sz. melléklet, d) pont). A webhungarikumok köre ennek analógiájára a .hu doméntartományban publikált, továbbá a más doméntartományban megjelent, de magyar nyelven, magyar szerzőtől származó, illetőleg a magyar vonatkozású egyéb nyelvű tartalom. A teljes magyar webtér feltárása még a legszűkebb értelmezés szerint is szinte lehetetlen feladat, a web változékonysága, az exponenciális növekedése miatt még a .hu doméntartomány milliós nagyságrendű webhelyeit is nehéz számba venni és nyilvántartani, és még nehezebb feladat a nem .hu alá regisztrált, illetőleg nem magyar nyelvű, de magyar vonatkozású tartalom feltárása. A .hu tartományba regisztrált webhelyek nyilvántartása különböző forrásokból (l. alább) megoldható, ám még ennek a szűkebb értelmezés szerinti webhungarikum-tartalomnak a feltárása esetén is csak szelektálva, a teljességre törekvés nélkül lehet csak a kurátori feladatokat ellátni. A szelektálás egyfelől kényszerű: a teljes doméntartomány tartalmának csak egy része érhető el nyilvánosan, másfelől bizonyos tartalmakat ki kell zárnunk a gyűjtőkörből (pl. pornográf vagy káros, rosszindulatú, adathalász, vagy más módon félrevezető tartalmak).

A webtéraratás jogszabályi keretei

2020. május 19-én megszavazta az országgyűlés a kulturális törvényt módosító javaslatokat, köztük azt is, amely a nemzeti könyvtár feladatává teszi a webtartalom megőrzését (2020. évi XXXII. törvény, 7. A muzeális intézményekről, a nyilvános könyvtári ellátásról és a közművelődésről szóló 1997. évi CXL. törvény módosítása). Az 1. §-ba bekerült a g) és h) pont: a törvény célja

"g) gondoskodni a digitális formában, interneten szolgáltatott és nyilvánosságra hozott tartalom (a továbbiakban: webtartalom) archiválásáról, használatának biztosításáról, valamint

h) szabályozni a webtartalom archiválásával kapcsolatos feladatok ellátását, az archivált webtartalmak használatának lehetőségeit." (XXXII. törvény)

Az 1997. évi CXL. törvénybe bekerültek a webarchiválásról szóló részek (elsősorban az 59/A §):

"(1) A webtartalom archiválás keretében, webaratással történő begyűjtésének, feldolgozásának, másolásának, hosszú távú megőrzésének és webarchívumba rendezésének, továbbá felhasználásának feladatát (a továbbiakban együtt: webarchiválás) a nemzeti könyvtár látja el. A könyvtárak együttműködnek a nemzeti könyvtárral az archivált webtartalom hozzáférhetővé tételében" (Kult. tv.).

A törvény előírja továbbá a tartalomgazdák együttműködési kötelezettségét, valamint szabályozza a nemzeti könyvtár webarchiválással összefüggő adatkezelési jogait.

A Kormány 626/2020. (XII. 22.) számú rendeletben kiadta a webarchiválás részletes szabályait (Korm. rend.). Ez utóbbi értelmében a nemzeti könyvtár évente kétszer köteles webtérszintű aratást végezni. A rendelet szabályozza továbbá az archivált tartalom használatának a feltételeit és lehetőségeit is.

A webtéraratás kivitelezése, szelekció, seedlista

A webtér bármely értelmezését vesszük, a teljes körű és minden tartalmat mentő aratás lehetetlen, legfeljebb viszonylagos teljességre törekvő lehet. Teljességre törekvő webaratást azok a webarchívumok végeznek, amelyek gyűjtőkörébe relatíve kisebb méretű tartalom tartozik, pl. a Luxemburgi Nemzeti Könyvtár, az Osztrák Nemzeti Könyvtár, a Cseh Nemzeti Könyvtár (vö. Brakker & Kuybishev 2013; Böhm 2014), de a teljességre törekvés itt is csak a nyilvánosan elérhető webhelyek bizonyos mélységig történő mentését jelenti. Az esetek többségében valamilyen szempontrendszer szerinti szelekció, illetve szűkítés történik (pl. bizonyos tartalmak kizárása gyűjtőkör alapján, bizonyos fájltípusok kizárása, az aratás mélységének és/vagy a letöltött tartalom méretének a korlátozása stb.). A nemzeti könyvtárak feladata e tekintetben az, hogy megpróbálják biztosítani a webes tartalom megfelelő reprezentativitását az archívumban, hogy legalább pillanatfelvétel-jelleggel megőrződjenek (vö. még Drótos 2017).

A magyar webarchívum webtérseedlistája

A webtéraratás egyik fontos feltétele egy megfelelő, URL-eket tartalmazó címlista, ún. seedlista összeállítása. A seed URL kiindulópont egy crawler számára egy URL, amelyet elsőként arat le, és utána követi az abban található linkeket. Az URL rendszerint egy webhely kezdőlapja vagy egy olyan weboldal, ahonnan sok link mutat a saját szerverre vagy más szerverekre. Nagyobb méretű aratásoknál a crawler egy seedlistát kap, és abból indul el több szálon egyszerre, amely lista nagyon sok (akár több millió) URL-t is tartalmazhat. Egy jó seedlista összeállítása és karbantartása fontos feltétele az adott webarchiválási cél elérésének.

A magyar webarchívum indulásakor különböző webes forrásokból gyűjtötte az URL-eket (linkgyűjtemények, tematikus gyűjtőoldalak stb., Drótos & Németh 2018). A különgyűjteményekhez manuális kurátori munkával gyűjtött címek is bekerültek a webtérseedlistába, amely a saját aratásokból kigyűjtött új URL-ekkel tovább bővül minden évben, továbbá egy ajánló űrlapon keresztül bárki küldhet be archiválásra javasolt webhelyet (https://forms.office.com/r/Cz3GsHSQEd). Nagy előrelépést jelentett az Internet Archive-val kötött megállapodás, amelynek keretében megkaptuk az általuk ismert .hu doméntartományba tartozó URL-eket 2022-ben (több mint 5 millió).[1] A fenti források alapján a nemzeti könyvtár által feltárt hungarikumnak tekinthető webhelyek száma több mint 6,5 millió. Ezekből az évente kétszer elvégzett aratások során változó számú cím került aratásra: a nem gyűjtőkörbe tartozó vagy inaktív, megszűnt webhelyek értelemszerűen nem kerülnek a seedlistába.

Aratás kezdete	Aratás vége	Kiinduló URL-ek száma	Letöltött URL-ek száma	Letöltött tartalom	Eltárolt új tartalom (tömörítés nélkül)
2024-12-19	2025-01-15	817 542	111 908 302	7 906 GB	3 955 GB
2024-06-24	2024-07-23	865 982	129 728 757	8 788 GB	5 453 GB
2024-01-11	2024-02-03	1 371 617	138 409 426	8 766 GB	6 531 GB
2023-10-04	2023-10-31	992 303	126 850 047	7 830 GB	6 451 GB
2022-12-02	2022-12-20	1 371 617	158 416 570	8 261 GB	6 687 GB
2022-06-24	2022-07-20	1 371 617	174 282 398	8 891 GB	6 132 GB
2021-12-26	2022-01-03	433 863	69 356 724	5 372 GB	2 414 GB
2021-07-07	2021-07-12	433 863	71 878 955	5 495 GB	3 128 GB
2020-12-30	2021-01-04	251 230	47 881 581	4 140 GB	2 300 GB
2020-06-30	2020-07-05	269 430	46 380 598	3 608 GB	2 400 GB
2019-12-23	2020-01-02	246 819	110 367 190	6 387 GB	6 387 GB
2018-09-24	2018-09-28	291 078	172 639 350	10 287 GB	9 138 GB

1. ábra: A magyar webarchívum webtéraratásainak főbb statisztikai adatai

Az aratás időtartamát a kiinduló URL-ek száma és a várhatóan letölthető új tartalom becsült mennyisége alapján álltjuk be. Előfordul, hogy technikai fennakadások, hibák miatt újra kell indítani, ilyenkor a teljes futásidő hosszabb. A kiinduló URL-ek száma 2021-ig lassabban növekedett, a 2022-es jelentősebb növekményt az Internet Archive-val megkezdett együttműködésünknek köszönhettük. Ezután 2024 nyarán újraellenőriztük a seedlistát, és elég nagy számban kellett eltávolítanunk URL-eket, mert közben a webhelyek elég jelentős része inaktívvá vált. (A 2023-as kisebb szám magyarázata az, hogy a három részaratásból az egyik meghiúsult technikai problémák miatt.) A letöltött URL-ek számát és a letöltött tartalom méretét alapvetően a bejárt webhelyek tartalma határozza meg (hány linket tud követni a crawler, mennyi új tartalom került fel a webhelyekre stb.)

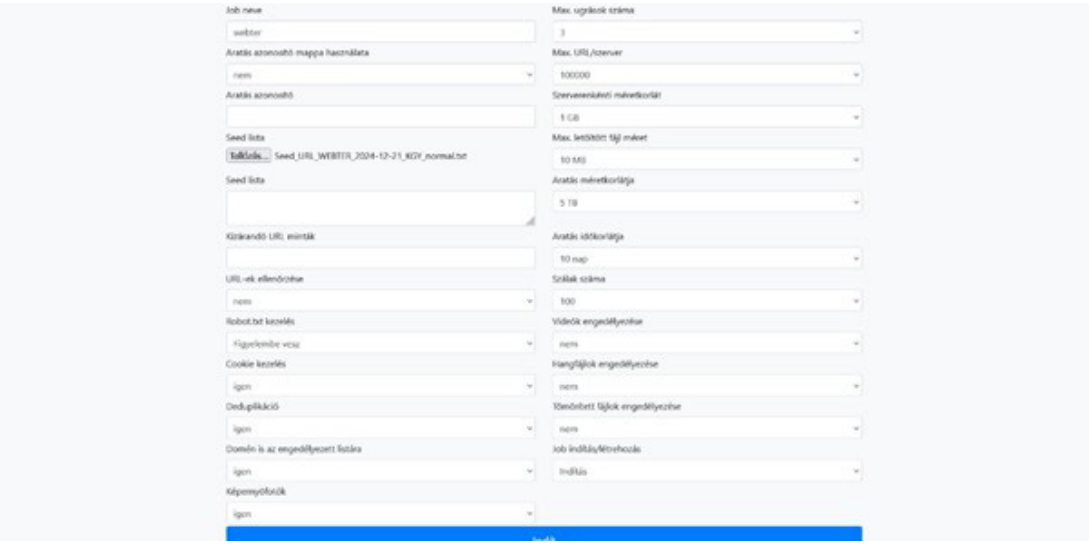
1 Folyamatban van egy másik nemzetközi webarchiváló szervezet, a CommonCrawl (https://commoncrawl.org/) seed URL-jeinek a feldolgozása a host index felhasználásával, amelyben a webhely nyelve szerint is lehet keresni, ez jelentős mértékben bővíti majd a nem .hu domántartományba regisztrált magyar nyelvű webhelyek nyilvántartását.

A magyar webarchívum aratórobotja és beállításai

A magyar webarchívum a webtéraratáshoz az Internet Archive 2003 óta fejlesztett aratórobotját, a Heritrixet használja (https://github.com/internetarchive/heritrix3). A szoftver Java nyelven íródott, Linuxon fut. Korábban az Internet Archive ARC-formátuma (https://archive.org/web/researcher/ArcFileFormat.php) volt a kimenete, később áttért a WARC-formátumra (https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1-annotated/), amely ISO 28500-as sorszámmal nemzetközi szabvány az International Internet Preservation Consortium gondozásában. Az aratórobot széleskörűen konfigurálható, beállítható többek között az ugrások száma, a letöltött tartalom mérete, a futásidő, és még sok minden más. Egy-egy aratáshoz (ún. jobhoz) külön seedlista tartozik. A Heritrix támogatja a deduplikációt, a letöltött tartalmat hash-értékkel látja el, amelyet adatbázisban tárol, és ha újra találkozik az adott tartalommal, akkor nem tölti le még egyszer.

A magyar webtér aratásához három jobot használunk. Külön beállításokkal futtatjuk az ún. tömeges aldoméneket tartalmazó webhelyek mentését (mint amilyenek pl. a lap.hu oldalak), mert ezek általában üzleti- vagy reklámcélú felületek, sokszor kevésbé összetett struktúrával, valamint kevésbé értékes tartalommal, ezért pl. kisebb mélységet elegendő beállítani az esetükben a linkek követéséhez. Ugyancsak külön kell kezelni azokat a webhelyeket, amelyek nem használnak ún. robots.txt fájlt. Ez egy egyszerű szövegfájl a gyökérkönyvtárban, mellyel a webhely adminisztrátora szabályozni tudja, hogy a keresőgépek és az archiváló szolgáltatások által indított crawlerek a webszerveren levő tartalom mely részét járhatják be, sőt akár ki is tilthatja őket teljesen. A tiltások és engedélyek alkönyvtárakra, fájlokra és az egyes crawlerekre (user-agentekre) korlátozhatók. Mivel a magyar jogszabályok szerint a nemzeti könyvtár akkor is jogosult az archiválásra, ha egy webhelyen nincs robots.txt, ezért külön jobként aratjuk az ilyeneket, ignorálva a robots.txt-t, amelyet a Heritrix egyébként alapértelmezetten figyelembe vesz. A többi webhely kerül a harmadik, ún. normál seedlistába.

A jobok konfigurációját a crawler-beans.cxml fájl tartalmazza, amely közvetlenül is szerkeszthető a Heritrix böngészőben futtatható felületén (az ún. engine-en keresztül). Az aratások konfigurálásának megkönnyítésére egy Java nyelven írt webalkalmazást használunk (Kaptafa), amely a beállításokat beleírja a crawler-beansbe.



2. ábra: A 2024. évi második webtératás konfigurációja a Kaptafában

A 2. ábrán látható, hogy 2024 telén milyen beállításokkal indítottuk el a normál webtérseedlista aratását (a legfontosabbakat kiemelve): 3 ugrást engedélyeztünk a robotnak (a linkek követésében), egy-egy szerverről legfeljebb 100 ezer URL meglátogatását, legfeljebb 1 GB tartalom letöltését, 10 MB-nál nagyobb fájlokat ne töltsön le, nem tölthet le hang- és videofájlokat, a teljes aratás méretkorlátja 5 TB, és legfeljebb 10 napig futhat a job. A 2024. évi második webtéraratás három jobja esetében a seed URL-ek a száma 817 542 volt, a robot összesen majdnem 112 millió URL-t látogatott meg, a letöltött új tartalom mérete közel 4 TB volt.

A webtérseedlista karbantartása

Ahhoz, hogy webtéraratások a lehető legjobb minőségűek legyenek, a webtérseedlistát rendszeresen karban kell tartani. Ez az alábbi kurátori feladatokból áll.

Rendszeresen összegyűjtjük a korábbi mentéseinkből az új, addig ismeretlen URL-eket. Ehhez az archívum WARC-fájljaiból kell kigyűjtenünk az URL-eket, és össze kell őket vetnünk a már ismert címek listájával. A WARC-fájlok teljes átfésülése többéves tapasztalataink alapján leghatékonyabban Python-scriptekkel lehetséges, a korábbi scriptjeink, amelyek a bináris fájlokban reguláris kifejezésekkel történő mintaillesztéssel dolgoztak, sokkal lassabbnak és kevésbé hatékonynak bizonyultak, annak ellenére, hogy kizárólag Linux-parancsokat használtak. A jelenlegi megoldásunk a warcio.archiveiterator Python-könyvtárra épül, amely képes célzottan elérni és kinyerni adatokat a WARC-fájlokból.

A sebesség további növelése érdekében a multiprocessing Python-könyvtár segítségével többszálúsítjuk a folyamatot. A script képes az inkrementális lekérdezésre, a bevezetésétől kezdve csak az új WARC-állományokat fésüli át.

Ellenőrizni kell, hogy az ismert, valamint az újonnan kinyert URL-ekhez tartozó webhelyek működnek-e. Ehhez a webhelyek aktuális HTTP-státuskódját kérdezzük le. A legjobb módszer kialakítását ez esetben is hosszas tesztelés előzte meg, amelynek eredményeképpen itt is a Python tűnik a legjobb megoldásnak az asyncio és httpx könyvtárak használatával, párhuzamosan többszálúsítva. Az URL-eket HTTP és HTTPS protokollokkal, valamint www. aldoménnel és anélkül is le kell kérdeznünk, mert előfordul, hogy egy-egy webhely csak az egyik protokollal, vagy csak www. aldoménen működik, továbbá az is, hogy új doménre átirányítás csak valamely kombinációról történik meg. A státuskódok lekérdezésekor rögzítjük azokat az átirányításokat, amelyek eltérő doménre történnek, mert így tudjuk a webhelyek alternatív elérési útjait nyilvántartani. A 400-as és 500-as státuskódú webhelyek kikerülnek az aktuális seedlistából.

A működőnek tűnő webhelyek (200-as vagy 300-as státuskódúak) esetében további klasszifikációs feladatok vannak. Az újak vagy megváltozott tartalmúak esetében el kell döntenünk, hogy gyűjtőkörbe tartoznak-e. Minden esetben felül kell vizsgálnunk, hogy a HTTP-státuskód valóban tükrözi-e a webhely állapotát: nagyon gyakori eset, hogy a 200-as kód alapján működőnek látszik, de valójában egy ún. parkoló domén, azaz nincs mögötte valódi tartalom, csak a megvásárolhatóságáról szóló információ. Az is gyakori, hogy az elérhetetlenné váló tartalmak helyére egy olyan weboldal kerül, amelyen nincs valódi tartalom, csak tájékoztató szöveg (pl. "Az oldal már nem található."), vagy egyszerűen üres a HTML. A nem gyűjtőkörbe tartozó vagy valójában érdemi tartalmat nem szolgáltató webhelyeket el kell távolítanunk az aktuális seedlistából. Mindezek kiszűréséhez, valamint a webhelyek további klasszifikációjához lekérdezzük a HTML head tag title címkéjének az adatát. Ez az esetek nagy részében a fenti és további osztályozási feladatokhoz jó kiindulópontként szolgál. Ez a lekérdezés a legösszetettebb feladat, amelyet ugyancsak Pythonnal lehet a leghatékonyabban és leggyorsabban elvégezni. A script első lépésként az aiohttp, asyncio és a Beautifulsoup könyvtárak segítségével párhuzamosan lekérdezi a HTML-tartalmat minden kombinációban (HTTP és HTTPS protokollal, www-vel és anélkül), követve az átirányításokat, és az első esetben, amikor sikeresen ki tudja venni a title tag tartalmát, az adatot a kimenetbe írja, majd továbblép. Ha ezzel a módszerrel nem sikerül hozzájutni a title-adathoz (pl. gyakran előfordul dinamikusan generált tartalmak esetében), akkor a selenium Python-könyvtár használatával elindít egy Chromiumot headless módban, és úgy próbálja meg kinyerni a title-t. Annak ellenére, hogy a title időnként üres vagy értelmetlen adatot tartalmaz, a lekérdezés alapján mégis elég jól lehet klasszifikálni a webhelyeket a 200-as kód ellenére inaktív, nem gyűjtőkörbe tartozó és tömeges aldoménként

vagy normál doménként aratandó kategóriákba. Ezt a feladatot a jövőben tovább lehet fejleszteni felügyelt gépi tanulással, valamint a webhelyek szövegének az elemzésével.

Az aratások megkezdése előtt még azt is meg kell állapítanunk, hogy az aratandó webhelyen található-e robots.txt (l. fentebb). Ezt a feladatot is Python-script végzi. Az aratások utáni kurátori teendők közé tartoznak még a képernyőkép-készítés a mentett webhelyek kezdőoldaláról, a metaadatok és statisztikák előállítása, amely folyamatok nagy része is automatizáltan történik.

A nemzeti könyvtár több éves tapasztalatot szerzett a magyar webtér aratása során, és folyamatosan fejleszti a tevékenységét annak érdekében, hogy az internetes tartalmak mint a kulturális örökségünk részei megőrződjenek az utókor számára. Az OSZK az International Internet Preservation Consortium tagjaként igyekszik a nemzetközi közösség tapasztalatait és fejlesztéseit is beépíteni a munkafolyamataiba. A mentések az Országos Széchényi Könyvtár olvasótermének dedikált gépein kereshetőek, böngészhetőek, kutathatóak párhuzamosan az Internet Archive által mentett tartalmakkal. A magyar webtér aratása természetesen tovább fejleszthető, folyamatban van egy adatbázis kialakítása, amely nem csupán a törvény által is előírt nyilvántartó funkciót szolgálja majd, hanem a folyamataink automatizálásának alapjaként is funkcionál, törekszünk továbbá a webhungarikumok feltárásának a kiszélesítésére is, hogy a .hu doméntartományba tartozó tartalmak mellett más magyar vonatkozású webhelyek is bekerüljenek az archívumba.

Irodalom

2020. évi XXXII. törvény a kulturális intézményekben foglalkoztatottak közalkalmazotti jogviszonyának átalakulásáról, valamint egyes kulturális tárgyú törvények módosításáról. https://net.jogtar.hu/jogszabaly?docid=a2000032.tv Letöltve: 2025.03.25.

Böhm, T. (2014). Development of the National Library of the Czech Republic 2011–2016: Past, Present and Future. Alexandria: The Journal of National and International Library and Information Issues, 25(3), 17–24. https://doi.org/10.7227/ALX.0028

Brakker, N. V., & Kujbyshev, L. A. (2013). The Experience of the National Libraries Abroad of the Collection and Longterm Preservation of Internet Resources. *Bibliotekovedenie* [*Library and Information Science* (*Russia*)], *2*, 88–96. https://doi.org/10.25281/0869-608X-2013-0-2-88-96

Drótos, L. (2017). Az internet archiválása mint könyvtári feladat. *Tudományos és Műszaki Tájékoztatás*, *64*(7–8), 361–371.

Drótos, L., & Németh, M. (2018). Az OSZK-ban folyó kísérleti webarchiválási projekt első évének tapasztalatai. *Tudományos és Műszaki Tájékoztatás*, *65*(7–8), 389–400.

Korm. rend. 626/2020. (XII. 22.) Korm. rendelet a webarchiválás részletes szabályairól. https://njt.hu/jogszabaly/2020-626-20-22 Letöltve: 2025.03.25.

Kult. tv. 1997. évi CXL. törvény a muzeális intézményekről, a nyilvános könyvtári ellátásról és a közművelődésről. https://net.jogtar.hu/jogszabaly?docid=99700140.tv Letöltve: 2025.03.25.