Metadata supported full-text search in a web archive

Márton Németh – László Drótos

National Széchényi Library
Hungary

DONATE

INTERNET ARCHIVE WayBackMachine

Explore more than 362 billion web pages saved over time

library

http://loc.gov/

*the **library** of congress*

capture(s) from **1996** to **2017** | Site stats

http://dpls.lib.or.us/

*public **library***

capture(s) from **1999** to **2011** | Site stats

http://ntl.bts.gov/

library

capture(s) from **1998** to **2017** | Site stats

http://muse.mcgill.ca/

library

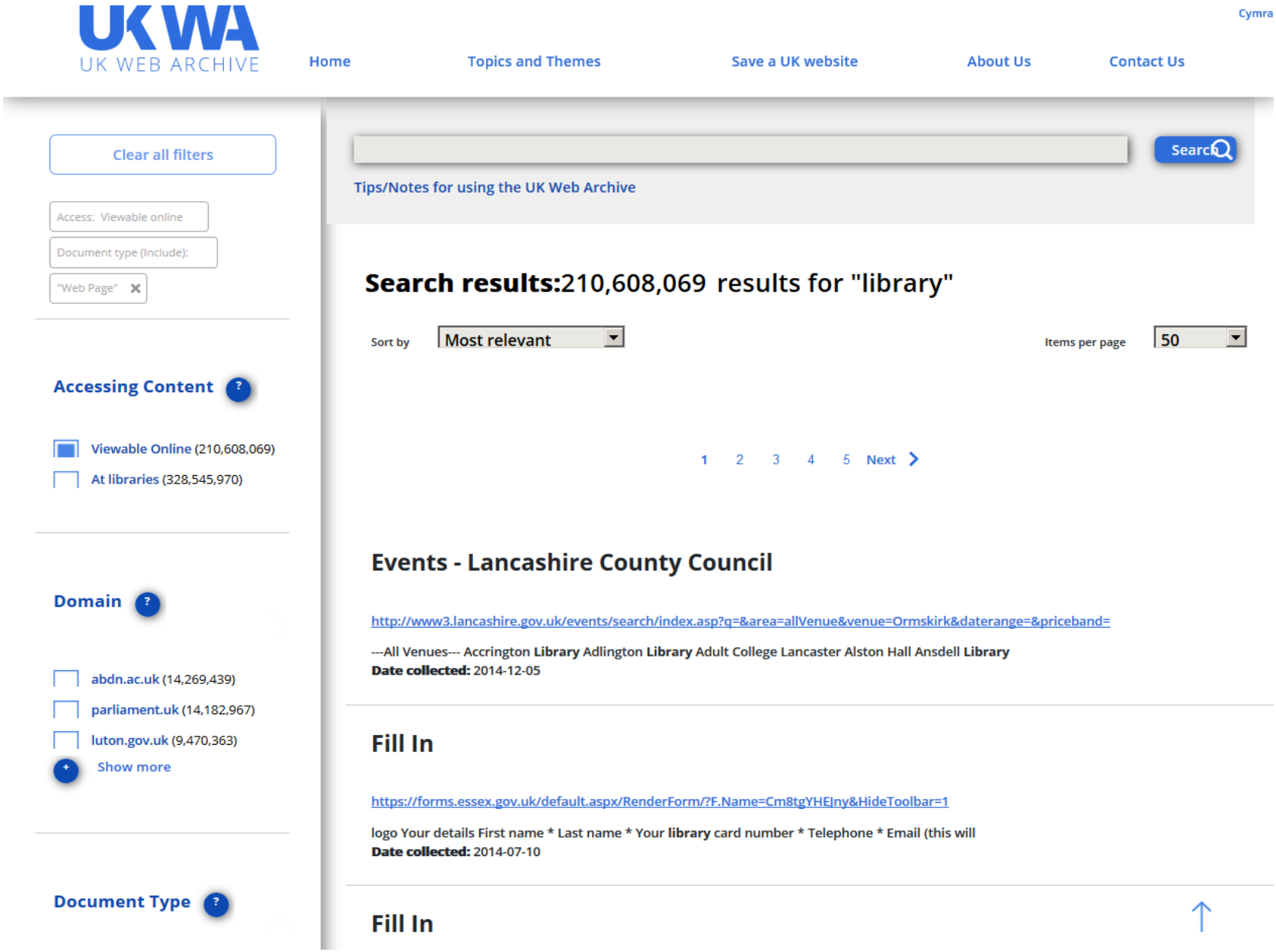
capture(s) from **1998** to **2006** | Site stats

http://orl.bc.ca/

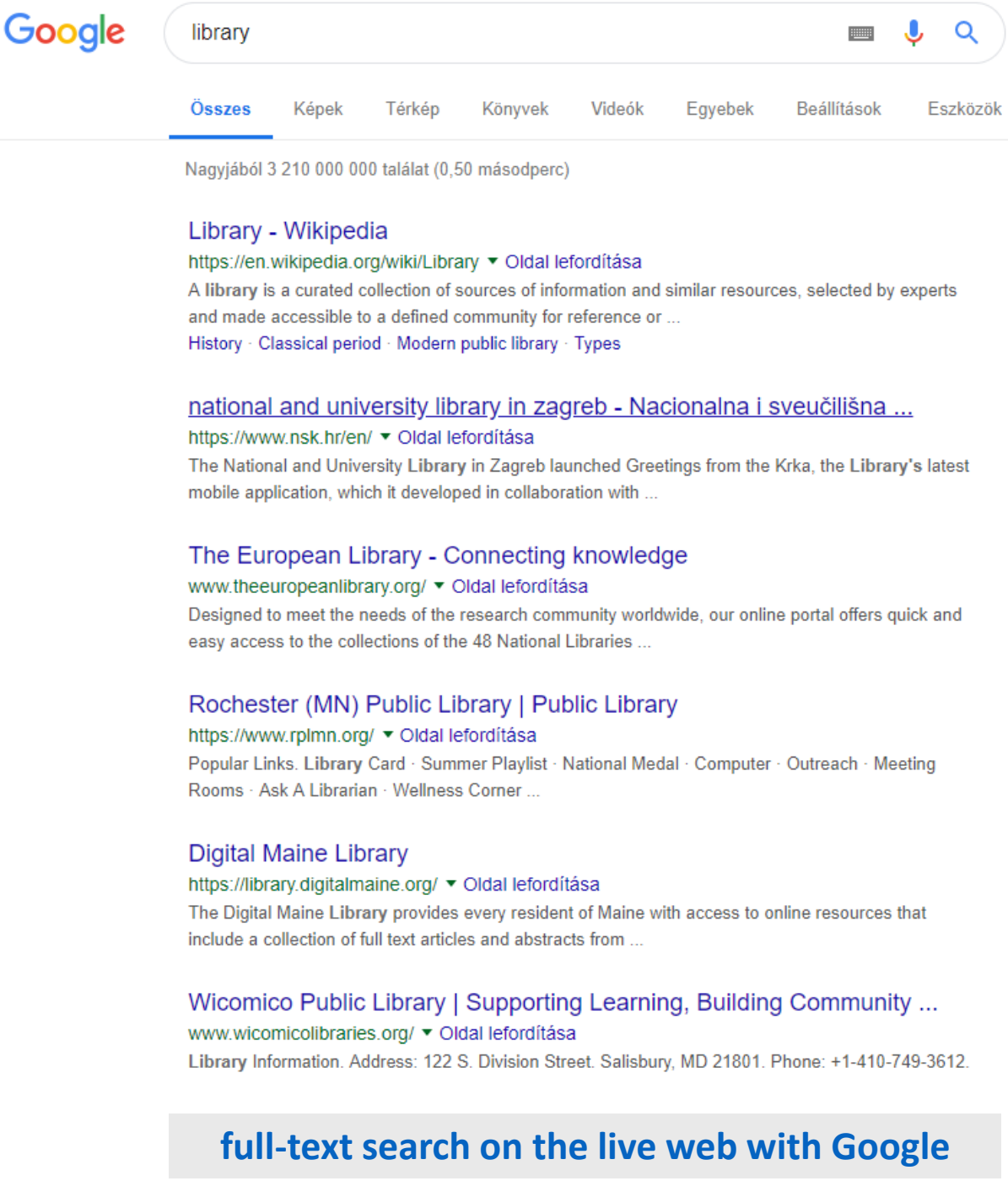
library

capture(s) from **2000** to **2017** | Site stats

search by web site names in the Wayback Machine



full-text search in the UK Web Archive



full-text search on the live web with Google

Main challenges for web archives:

- lack of full-text search capability (e.g. because of the sheer size of the archive, or because GDPR considerations);
- lack of relevance (how to get relevant hits from a huge set of archived webpages full of partially duplicated content and „garbage”?);
- such a knowledge base, that Google has, is out of question for the web archiving community (over 200 ranking factors: geo targeting, user browsing and search history, freshness of the content, number of organic clicks, artificial intelligence etc.)

What can we do instead?

- metadata supported full-text search
- semantic web-based query tools and services

Metadata types (on website level):

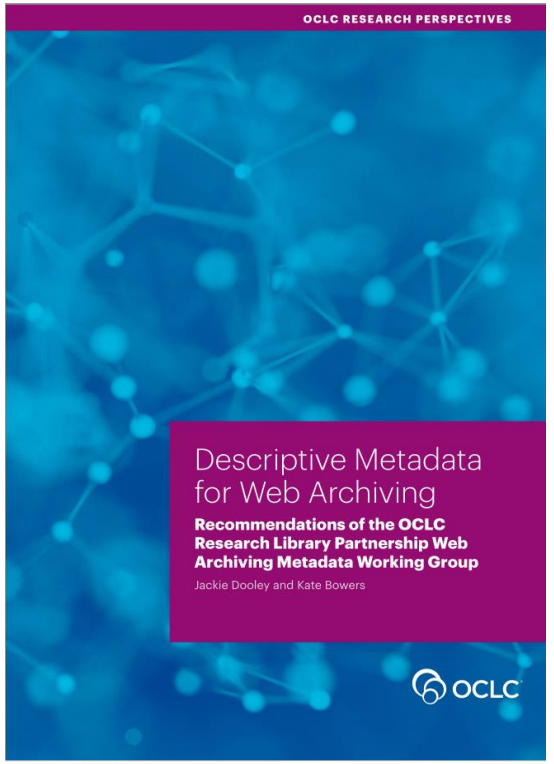
- bibliographic:
  e.g. title *(lots of variations)*, creator/contributor/publisher *(uncertain roles)*, rights *(unclear legal status)*, dates *(what kind of dates?)*, subject/type *(very mixed content)* ...

- administrative:
  e.g. curator, nominator, urgency, permission request, harvesting schedule, quality assurance, access ...

- technical:
  e.g. original CMS, harvester software, harvest parameters, size of the downloaded content, storage, long-term preservation ...

Recommendations for metadata:

- ISO/TR 14873:2013 – Statistics and quality issues for web archiving (collection level indicators)
- Descriptive Metadata for Web Archiving
  OCLC Web Archiving Metadata Working Group
  (mostly site-level bibliographic data fields, based on the Dublin Core schema)
- Metadata Application Profile for Description of Websites with Archived Versions
  New York Art Resources Consortium
  (site-level, MARC/RDA)

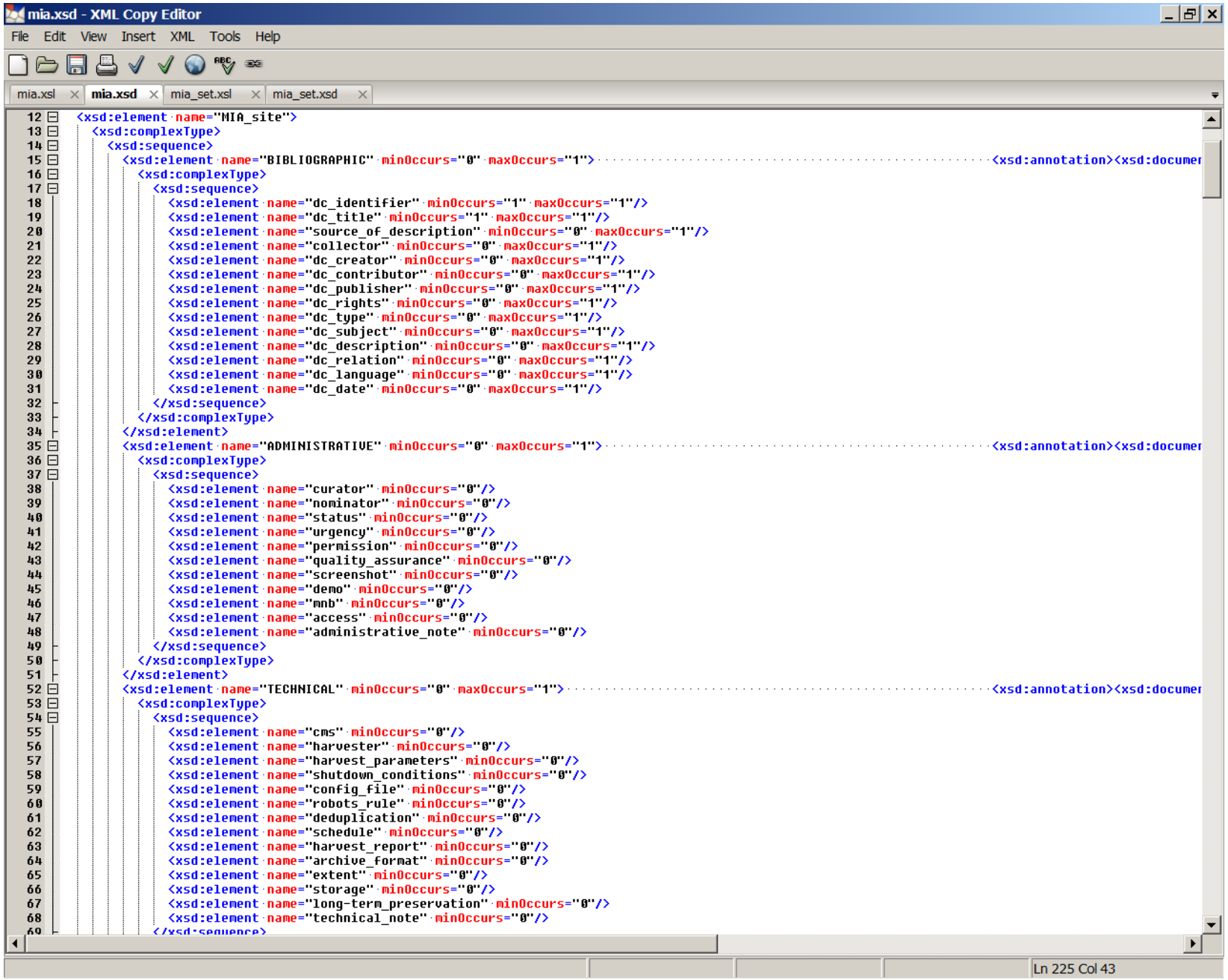


	A	B	C	D	E	F	G	H	I	J	K
1	Mezőkód	Mezőnév	Címke	XML tag	MARC 21 mező	Minősítők	Státusz	Adattípus	Ismételhető	Kötött szótár	Megjegyzés, példa
2	**EGYEDI WEBHELYEK**										
3	**LEÍRÓ METAADATOK**										
4	L01	**ID**	azonosító	mia_id	001	nincs	kötelező (automa	"MIA-" előtag, majd	nem	nincs	MIA-000123
5	L02	**Title**	cím	dc_title	245; 246; 730;740	egységesített cím (+besorc	kötelező (legalább	rövid szöveg	korlátlanul	nincs	A besorolási cím az egységesített cím
6	L03	**URL**	URL cím	dc_identifier	856 u almező	eredeti URL, alternatív ered	kötelező (legalább	URL	korlátlanul, de az eredeti	nincs	
7	L04	**Source of Description**	a leírás forrása	source_of_descri	588	forrás neve, forrás URL-je,	opcionális	rövid szöveg, URL,	korlátlanul csoportosan	nincs	
8	L05	**Collector**	begyűjtő	collector	583 h almező;710;852 a	intézmény, szervezeti egys	ajánlott	rövid szövegek	korlátlanul csoportosan	nincs	
9	L06	**Creator**	létrehozó	dc_creator	100;110;700,710	intézmény, szervezeti egys	ajánlott	rövid szöveg	korlátlanul	nincs	
10	L07	**Contributor**	közreműködő	dc_contributor	700	szerző, közreadó, szerkes	opcionális	rövid szöveg	korlátlanul	nincs	
11	L08	**Publisher**	kiadó	dc_publisher	260 a; 260 b	kiadó neve, kiadó e-mail cí	ajánlott	rövid szöveg, e-mail	korlátlanul csoportosan	nincs	
12	L09	**Rights**	jogok	dc_rights	506; 540; 542	tartalmi jogtulajdonos neve,	ajánlott	rövid szöveg, e-mail	korlátlanul csoportosan	CC licencek: ht	
13	L10	**Genre/Form**	műfaj/típus	dc_type	655	műfaj, típus, változékonysá	ajánlott	rövid szöveg	korlátlanul	mindháromhoz	Pl. "honlap", "blog", "periodika", "porta
14	L11	**Subject**	téma	dc_subject	600; 610; 611; 630; 647;	főtéma, téma, altéma, tárgy	ajánlott	rövid szöveg	korlátlanul	a téma és altér	Pl. "Kulturális források", "Oktatási forr
15	L12	**Description**	tartalmi leírás	dc_description	500; 505; 520;540; 541;	nincs	ajánlott	hosszú szöveg	korlátlanul	nincs	
16	L13	**Relation**	kapcsolat	dc_relation	580; 730; 773; 787	részgyűjtemény azonosító,	opcionális	részgyűjtemény és	korlátlanul csoportosan	nincs	
17	L14	**Language**	nyelv	dc_language	008 35-37 byteok; 041; 5	nyelvkód, megjegyzés	ajánlott	3-betűs kód, rövid s	korlátlanul	https://www.loc	
18	L15	**Date**	dátum	dc_date	008 07-14 byetok; 245 f	aratás kezdése, utolsó me	ajánlott	YYYY-MM-DD forma	korlátlanul	az archiválási g	Pl. "naponta többször", "naponta", "het
19											
20	**ADMINISZTRÁCIÓS METAADATOK**										
21	A01	**Curator**	felelős archiváló mun	curator		munkatárs neve, e-mail cím	ajánlott	rövid szöveg	korlátlanul	nincs	
22	A02	**Nominator**	javaslattevő	nominator		javaslattevő neve, javaslatte	opcionális	rövid szöveg, e-mail	korlátlanul csoportosan	nincs	Részben automatizálható a kitöltése a
23	A03	**Status**	feladat státusza	status		nincs	ajánlott	rövid szöveg	nem	saját szótár	Pl. "javasolt", "még nem mentett", "má
24	A04	**Urgency**	sürgősség	urgency		sürgősség státusza, megje	opcionális	Y/N logikai érték, rö	nem	nincs	
25	A05	**Permission**	engedélyezés	permission		engedélyezés státusza, en	opcionális	rövid szöveg, YYYY-	korlátlanul csoportosan	a státuszhoz s	Pl. "nem kértünk engedélyt", "kértünk
26	A06	**Quality assurance**	ellenőrzés	quality_assuranc		ellenőrzés státusza, ellenő	opcionális	rövid szöveg, YYYY-	korlátlanul csoportosan	a státuszhoz é	Pl. "nincs ellenőrizve", "ellenőrizve", "n
27	A07	**Screenshot**	oldalkép	screenshot		oldalkép státusza, oldalkép	ajánlott	rövid szöveg, YYYY-	korlátlanul	a státuszhoz s	Pl. "nincs oldalkép", "van oldalkép", "a
28	A08	**Demo set**	demóba bekerült	demo		demó státusz	opcionális	Y/N logikai érték	nem	nincs	A nyilvános demó gyűjteménybe beker
29	A09	**National bibliography**	MNB-be javasolt	mnb		MNB státusz	opcionális	Y/N logikai érték	nem	nincs	Az MNB-be való felvételre javasolt. Aut
30	A10	**Access**	hozzáférés	access		hozzáférés státus, hozzáfé	ajánlott	rövid szöveg, YYYY-	nem	a státuszhoz s	"nyilvánosan szolgáltatható", "zárt háló
31	A11	**Administrative note**	adminisztrációs meg	administrative_nc		nincs	opcionális	hosszú szöveg	korlátlanul	nincs	
32											
33	**TECHNIKAI METAADATOK**										
34	T01	**CMS**	tartalomkezelő rends	cms		nincs	ajánlott	rövid szöveg	nem	saját szótár	Pl."Drupal", "Wordpress", "Dspace", "N
35	T02	**Harvester**	aratószoftver	harvester		nincs	ajánlott	rövid szöveg	korlátlanul	saját szótár	Pl. "Heritrix 3.x", "Heritrix 1.0", "HTTrac
36	T03	**Harvest parameters**	aratási paraméterek	harvest_paramet		mélység, ugrások száma,	opcionális	rövid szöveg	korlátlanul	nincs	
37	T04	**Shutdown conditions**	leállítási feltételek	shutdown_condit		max. URL-ek száma, max.	opcionális	rövid szöveg	korlátlanul	nincs	
38	T05	**Config file**	konfigurációs fájl	config_file		nincs	opcionális	URI	korlátlanul	nincs	Ha egyedi config fájlra van szükség, ak
39	T06	**Robots rule**	robots.txt kezelés	robots_rule		nincs	opcionális	Y/N logikai érték	nem	nincs	Annak jelzésére, hogy a robots.txt meg
40	T07	**Deduplication**	deduplikáció	deduplication		deduplikáció státusza, meg	opcionális	Y/N logikai érték, rö	nem	nincs	A megjegyzésbe pl. beírható, hogy a d
41	T08	**Schedule**	mentések ütemezés	schedule		első mentési időpont, ismé	ajánlott	YYYY-MM-DD HH-M	korlátlanul	nincs	
42	T09	**Harvest report**	aratási napló	harvest_report		nincs	opcionális	URI	korlátlanul	nincs	Automatizálható a kitöltése a Heritrix r
43	T10	**Extent**	mennyiségi jellemző	extent	300;347	mentések száma, webhely	opcionális	számok, dátum	korlátlanul csoportosan	nincs	Részben automatizálható a kitöltése a
44	T11	**Archive format**	archív formátum	archive_format		nincs	ajánlott	rövid szöveg	korlátlanul	saját szótár	Részben automatizálható a kitöltése a P
45	T12	**Storage**	raktári hely	storage		nincs	ajánlott	rövid szöveg	korlátlanul	nincs	Részben automatizálható a kitöltése a
46	T13	**Long-term preservation**	hosszú távú megőrze	long-term_preser		nincs	opcionális	hosszú szöveg	korlátlanul	nincs	Részben automatizálható a kitöltése a
47	T14	**Technical note**	technikai megjegyzé	technical_note		nincs	opcionális	hosszú szöveg	korlátlanul	nincs	

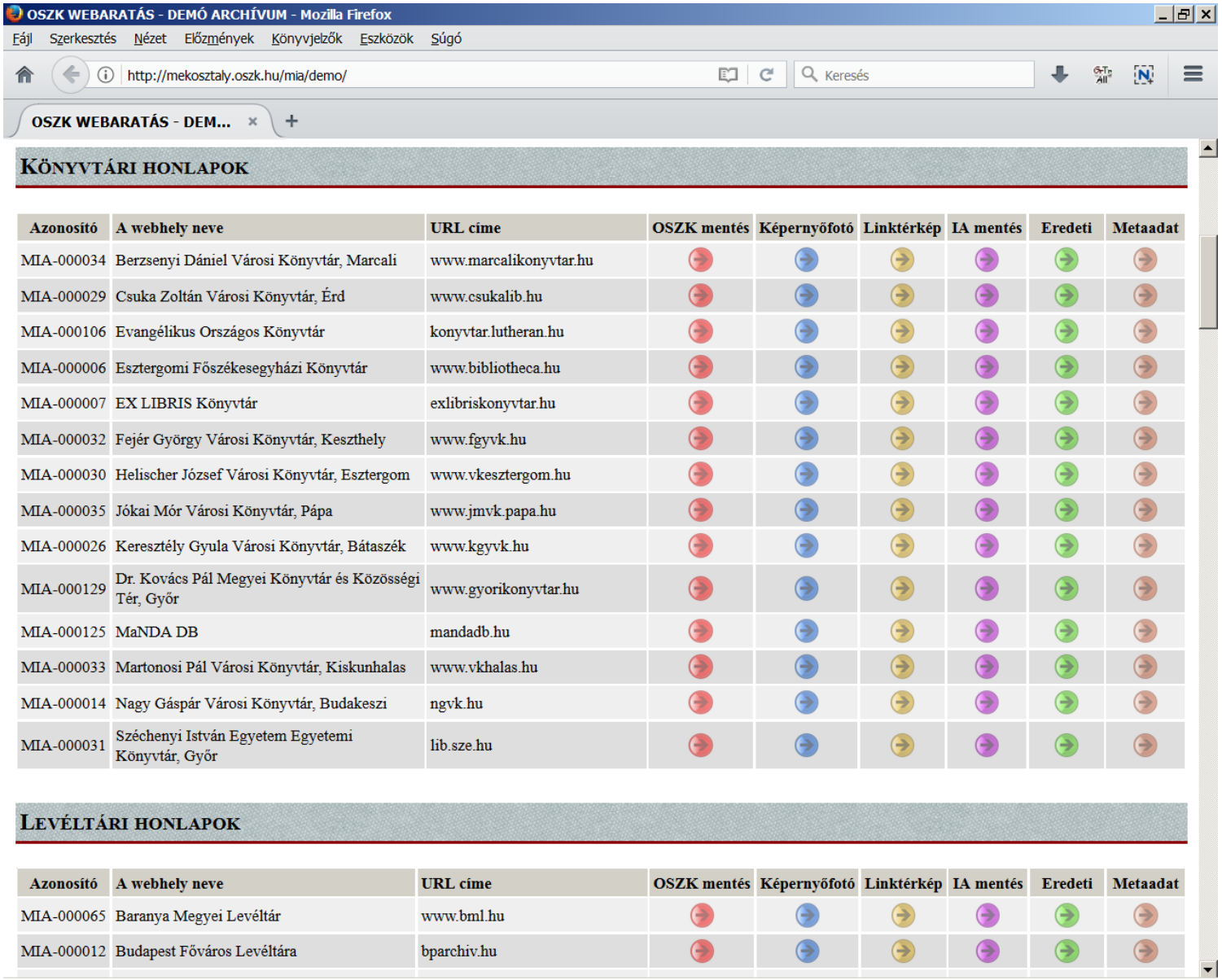
database plan for the Hungarian webarchive (website level)

Our metadata records:

- a small publicly available demo collection
- XSD (XML Schema Definition) and XSLT (Extensible Stylesheet Language Transformations) files
- predefined lists (e.g. genre, type, topic, subtopic, change frequency, harvest frequency, quality level)
- namespace links (person and geographic names)
- related sites (on the living web and in the archive)
- site-level and subcollection-level XML records
- manual data entry with XML Notepad (temporarily)



mia.xsd file (website level)



Könyvtári honlapok

Azonosító	A webhely neve	URL címe	OSZK mentés	Képernyőfotó	Linktérkép	IA mentés	Eredeti	Metaadat
MIA-000034	Berzsenyi Dániel Városi Könyvtár, Marcali	www.marcalikonyvtar.hu						
MIA-000029	Csuka Zoltán Városi Könyvtár, Érd	www.csukalib.hu						
MIA-000106	Evangélikus Országos Könyvtár	konyvtar.lutheran.hu						
MIA-000006	Esztergomi Főszékesegyházi Könyvtár	www.bibliotheca.hu						
MIA-000007	EX LIBRIS Könyvtár	exlibriskonyvtar.hu						
MIA-000032	Fejér György Városi Könyvtár, Keszthely	www.fgyvk.hu						
MIA-000030	Helischer József Városi Könyvtár, Esztergom	www.vkesztergom.hu						
MIA-000035	Jókai Mór Városi Könyvtár, Pápa	www.jmvk.papa.hu						
MIA-000026	Keresztély Gyula Városi Könyvtár, Bátaszék	www.kgyvk.hu						
MIA-000129	Dr. Kovács Pál Megyei Könyvtár és Közösségi Tér, Győr	www.gyorikonyvtar.hu						
MIA-000125	MaNDA DB	mandadb.hu						
MIA-000033	Martonosi Pál Városi Könyvtár, Kiskunhalas	www.vkhalas.hu						
MIA-000014	Nagy Gáspár Városi Könyvtár, Budakeszi	ngvk.hu						
MIA-000031	Széchenyi István Egyetem Egyetemi Könyvtár, Győr	lib.sze.hu						

Levéltári honlapok

Azonosító	A webhely neve	URL címe	OSZK mentés	Képernyőfotó	Linktérkép	IA mentés	Eredeti	Metaadat
MIA-000065	Baranya Megyei Levéltár	www.bml.hu						
MIA-000012	Budapest Főváros Levéltára	bparchiv.hu						

a snippet from the public demo collection webpage

icons: archived | screenshot | linkgraph | Internet Archive | original | metadata

```
<?xml version="1.0" encoding="utf-8"?>
<?xml-stylesheet type="text/xsl" href="http://mekosztaly.os
<MIA xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
/mia/xml/mia.xsd">
  <MIA_site>
    <BIBLIOGRAPHIC>
      <dc_identifier>
        <mia_id>MIA-000063</mia_id>
        <original_URL>http://obudaimuzeum.blogspot.hu</orig
        <seed_URL>http://obudaimuzeum.blogspot.hu/</seed_UR
        <archived_public_URL>http://193.6.201.202/owb/wayba
        <linkgraph_URL>http://193.6.201.202/solrwayback
/waybacklinkgraph.jsp?domain=obudaimuzeum.blogspot.hu</link
        <screenshot_URL>http://obudaimuzeum.blogspot.hu/</s
        <elsewhere_archived_URL>http://web.archive.org/web/
      </dc_identifier>
      <dc_title>
        <uniform_title>Óbudai Múzeum blogja</uniform_title>
        <original_main_title>Óbudai Múzeum - helytörté.net<
        <subtitle>Egy múzeum élete</subtitle>
      </dc_title>
      <source_of_description>
        <source_definition>Az Óbudai Múzeum blogja és a múz
        <source_URL>http://obudaimuzeum.blogspot.hu</source
        <source_URL>http://www.obudaimuzeum.hu</source_URL>
        <description_date>2018-07-05</description_date>
        <cataloger_name>Drótos László</cataloger_name>
      </source_of_description>
      <collector>
        <collector_organization>Országos Széchényi Könyvtár
        <collector_organization_unit>E-könyvtári Szolgáltat
        <collector_person>Visky Ákos László</collector_pers
        <project_name>MIA demó</project_name>
      </collector>
      <dc_creator>
        <creator_organization>Óbudai Múzeum</creator_organi
      </dc_creator>
      <dc_contributor></dc_contributor>
      <dc_publisher>
        <publisher_name>Óbudai Múzeum</publisher_name>
        <publisher_email>titkarsag@obudaimuzeum.hu</publish
        <publisher_URL>http://www.obudaimuzeum.hu</publishe
      </dc_publisher>
      <dc_rights>
        <content_rights_owner>Óbudai Múzeum</content_rights
        <service_rights_owner>Google Inc. : Blogger</servic
        <legal_status>jogvédett</legal_status>
        <contact_person>
          <contact_name>Népessy Noémi</contact_name>
          <contact_job>múzeumigazgató</contact_job>
          <contact_email>titkarsag@obudaimuzeum.hu</contact
        </contact_person>
      </dc_rights>
      <dc_type>
        <genre>Blog</genre>
        <owner_type>Intézményi</owner_type>
        <change_frequency>nem változó</change_frequency>
      </dc_type>
      <dc_subject>
        <main_topic>Kulturális források</main_topic>
        <main_topic>Tudományos források</main_topic>
        <main_topic>Oktatási források</main_topic>
        <topic>Köz- és magángyűjtemények</topic>
        <topic>Társadalomtudomány</topic>
        <topic>Ismeretterjesztés</topic>
        <subtopic>Múzeumok</subtopic>
```

ÓBUDAI MÚZEUM BLOGJA

MIA azonosító: MIA-000063
Eredeti URL: http://obudaimuzeum.blogspot.hu
Seed URL: http://obudaimuzeum.blogspot.hu/
Nyilvános OSZK-s archív URL: http://193.6.201.202/owb/wayback/*/http://obudaimuzeum.blogspot.hu/
Linktérkép URL: http://193.6.201.202/solrwayback/waybacklinkgraph.jsp?domain=obudaimuzeum.blogspot.hu
Oldalkép URL: http://obudaimuzeum.blogspot.hu/
Más nyilvános archív URL: http://web.archive.org/web/*/http://obudaimuzeum.blogspot.hu/

Egységesített cím: Óbudai Múzeum blogja
Eredeti főcím: Óbudai Múzeum - helytörté.net
Alcím: Egy múzeum élete

Leírás forrása: Az Óbudai Múzeum blogja és a múzeum honlapja
Forrás URL: http://obudaimuzeum.blogspot.hu
Forrás URL: http://www.obudaimuzeum.hu
Leírás dátuma: 2018-07-05
Leíró személy neve: Drótos László

Archiváló szervezet: Országos Széchényi Könyvtár
Archiváló szervezeti egység: E-könyvtári Szolgáltatások Osztálya
Archiváló személy: Visky Ákos László
Archiváló projekt: MIA demó

Létrehozó szervezet: Óbudai Múzeum

Kiadó neve: Óbudai Múzeum
Kiadó e-mail címe: titkarsag@obudaimuzeum.hu
Kiadó honlapja: http://www.obudaimuzeum.hu

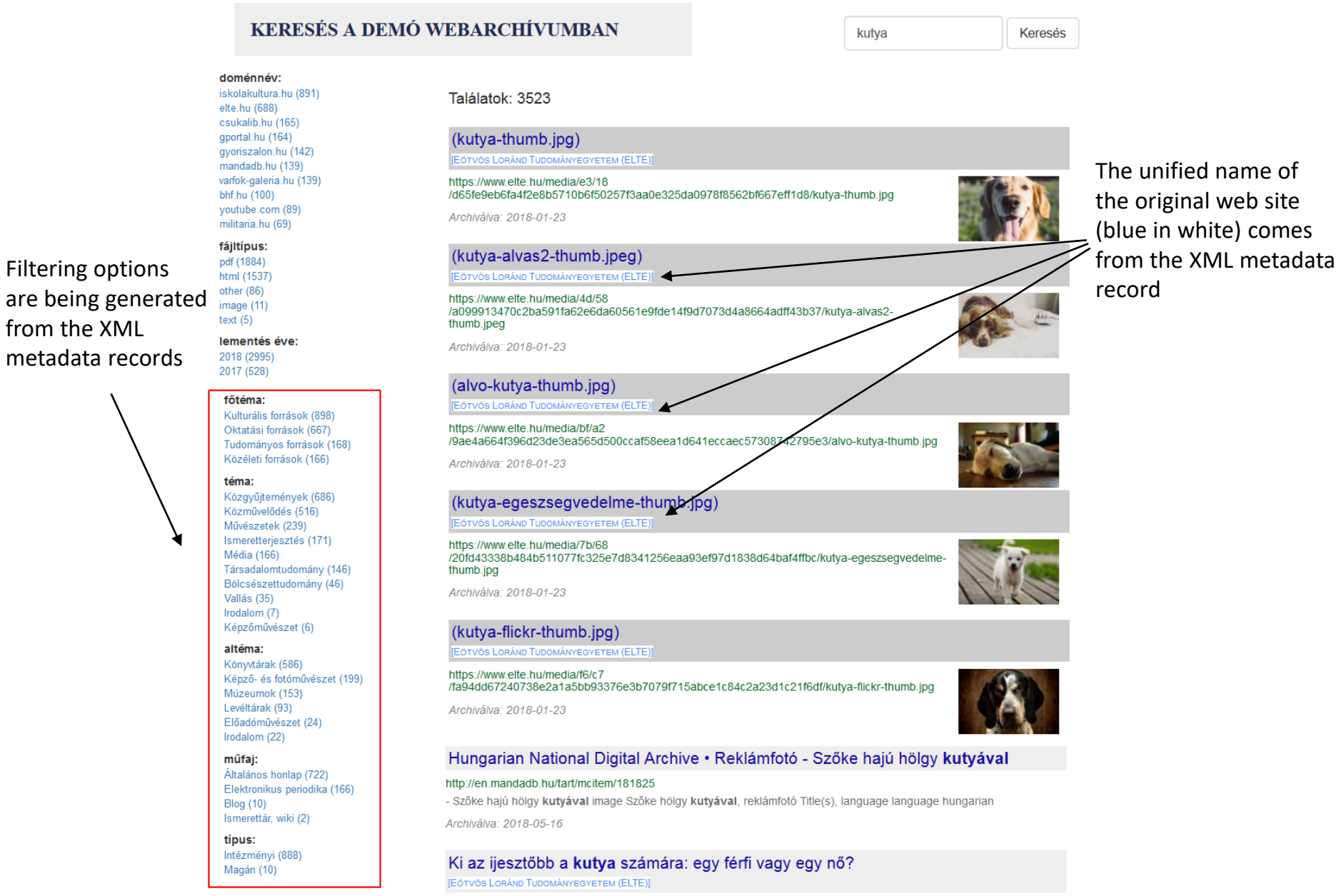
Tartalmi jogtulajdonos: Óbudai Múzeum
Szolgáltatási/technikai jogtulajdonos: Google Inc. : Blogger
Jogi státusz: jogvédett

Kapcsolati személy neve: Népessy Noémi
Kapcsolati személy beosztása: múzeumigazgató
Kapcsolati személy e-mail címe: titkarsag@obudaimuzeum.hu

Műfaj: Blog
Tulajdonosi típus: Intézményi
Változékonyság: nem változó

Fő témakör: Kulturális források
Fő témakör: Tudományos források
Fő témakör: Oktatási források
Témakör: Köz- és magángyűjtemények
Témakör: Társadalomtudomány
Témakör: Ismeretterjesztés
Altémakör: Múzeumok
Altémakör: Történelem, helytörténet
Tárgyszó: helytörténet
Tárgyszó: várostörténet

metadata of the Óbuda Museum's blog (original XML and converted HTML format)



Filtering options are being generated from the XML metadata records

The unified name of the original web site (blue in white) comes from the XML metadata record

filtering options by topic, subtopic, genre and type, (in-house developed Solr-based full-text search engine)

Future plans:

- database and form-based data entry interface (as part of the new nation-wide library system)
- cooperation with other memory institutions (e.g. shared cataloging)
- automatic and semi-automatic metadata generation (mostly technical and administrative data)
- incorporate metadata of important archived websites into the national bibliography
- faceted full-text hit lists filtered by metadata

Web archives and the semantic web:

- The absence of efficient and meaningful exploration methods of the archived content is a really major hurdle in the way to turn web archives to a usable and useful information resource.
- A major challenge in information science can be the adaptation of semantic web tools and methods to web archive environments.
- The web archives must be a part of the linked data universe with advanced query and integration capabilities, must be able to directly exploitable by other systems and tools.

International Journal on Digital Libraries
https://doi.org/10.1007/s00799-018-0251-0

CrossMark

Building and querying semantic layers for web archives (extended version)

Pavlos Fafalios[1] · Helge Holzmann[1] · Vaibhav Kasturia[1] · Wolfgang Nejdl[1]

Received: 15 September 2017 / Revised: 9 December 2017 / Accepted: 28 June 2018
© Springer-Verlag GmbH Germany, part of Springer Nature 2018

Abstract
Web archiving is the process of collecting portions of the Web to ensure that the information is preserved for future exploitation. However, despite the increasing number of web archives worldwide, the absence of efficient and meaningful exploration methods still remains a major hurdle in the way of turning them into a usable and useful information source. In this paper, we focus on this problem and propose an RDF/S model and a distributed framework for building semantic profiles ("layers") that describe semantic information about the contents of web archives. A semantic layer allows describing metadata information about the archived documents, annotating them with useful semantic information (like entities, concepts, and events), and publishing all these data on the Web as Linked Data. Such structured repositories offer advanced query and integration capabilities, and make web archives directly exploitable by other systems and tools. To demonstrate their query capabilities, we build and query semantic layers for three different types of web archives. An experimental evaluation showed that a semantic layer can answer information needs that existing keyword-based systems are not able to sufficiently satisfy.

Keywords Web archives · Semantic layer · Profiling · Linked data · Exploratory search

1 Introduction

Significant parts of our cultural heritage are produced and consumed on the Web. However, the ephemeral nature of the Web makes most of its information unavailable and lost after a short period of time. Aiming to avoid losing important historical information, a web archive captures portions of the Web to ensure the information is preserved for future researchers, historians, and interested parties in general.

This is an extended version of the paper: *P. Fafalios, H. Holzmann, V. Kasturia, & W. Nejdl, "Building and Querying Semantic Layers for Web Archives", 2017 ACM/IEEE-CS Joint Conference on Digital Libraries, June 2017.*

✉ Pavlos Fafalios
fafalios@L3S.de

Helge Holzmann
holzmann@L3S.de

Vaibhav Kasturia
kasturia@L3S.de

Wolfgang Nejdl
nejdl@L3S.de

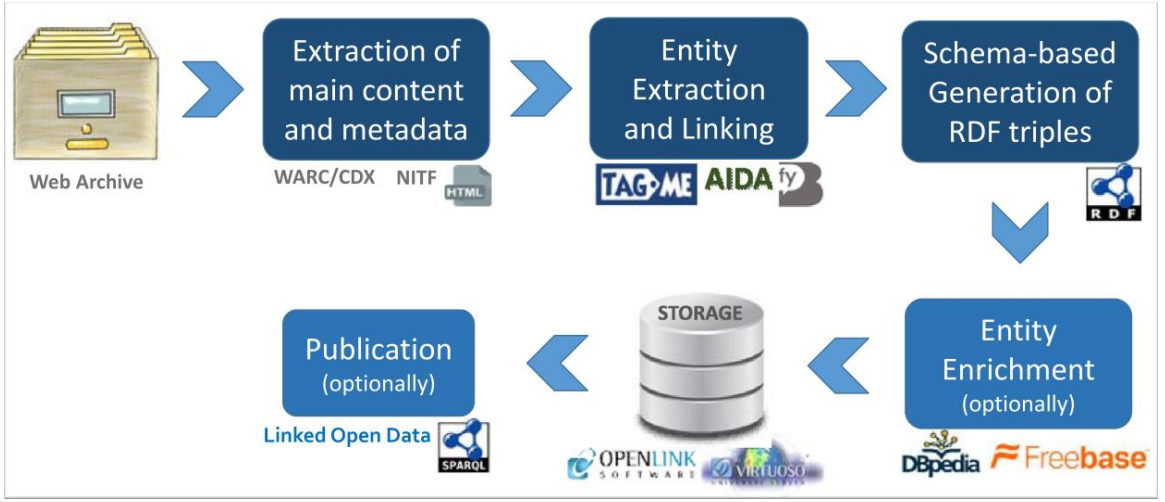
[1] L3S Research Center, Leibniz University of Hannover, Appelstr. 9a, 30167 Hannover, Germany

Despite the increasing number of web archives worldwide, the absence of efficient and meaningful exploration methods still remains a major hurdle in the way of turning web archives into a usable and useful source of information. The main functionalities offered by existing systems are to find older versions of a specific web page, to search on specific collections, or to search using keywords and filter the retrieved results by selecting some basic metadata values. However, for a bit more complex information needs, which is usually the case when exploring web archives, keyword-based search leads to ineffective interactions and poor results [48]. This is true especially for *exploratory search* needs where searchers are often unfamiliar with the domain of their goals, unsure about the ways to achieve their goals, or need to learn about the topic in order to understand how to achieve their goals [31]. Thus, for exploring web archives, there is the need to go beyond keyword-based search and support more advanced information seeking strategies [24,48,49].

To cope with this problem, we propose building semantic profiles ("layers") that describe semantic information about the contents of archived documents. Specifically, we base upon Semantic Web technologies and propose an RDF/S [11] data model that allows: (a) describing useful metadata information about each archived document, (b) annotating

Possible methods:

- extracting entities;
- generation of RDF triples;
- enrichment of entities from external sources;
- publication of linked data;
- advanced queries and ranking models based on semantic data.



The process of constructing a semantic layer in the Open Web Archive data model, proposed by Fafalios, Holzmann, et al. in 2018.

Future plans:

- automatic entity identification and extraction from the full text (e.g. names, events, concepts)
- enriching metadata from external sources (e.g. DBpedia)
- combining with the national namespace of the NSZL (in beta phase)
- metadata enrichment from namespaces and thesauri
- elaborate a PhD project (partners needed)

Thank you for your attention! Questions?

- Hungarian web archiving project: http://mekosztaly.oszk.hu/mia/
- Demo web archive: http://mekosztaly.oszk.hu/mia/demo/
- Selected bibliography on web archiving: http://mekosztaly.oszk.hu/mia/doc/webarchivalas-irodalom.html
- Contact e-mail: mia@mek.oszk.hu