

ARCHIVING OF WEBSITES AND BORN-DIGITAL DOCUMENTS IN SLOVAKIA

Jana Matúšková, Peter Hausleitner, Andrej Bizík

University Library in Bratislava

8th Colloquium of Library and Information Experts of the V4+ Countries

19.6.2019



NATIONAL PROJECT DIGITAL RESOURCES

- ▶ web archiving in Slovakia - first attempts 2005/2006 - ULIB participated in project Web Cultural Heritage
- ▶ there were no needed legislative, organisational and technical conditions for systematic harvesting and archiving of web pages and original electronic resources in Slovakia
- ▶ 2015 – University Library in Bratislava - national project
- ▶ Digital Resources – Web Harvesting and e-Born Content Archiving
- ▶ implementation: 01. 04. 2015 – 31. 12. 2015 (application development, ICT installation, pilot harvests)

DIGITAL RESOURCES - ORGANISATIONAL SUPPORT

Digital Resources Project - Mission & Goal:

- ▶ to create the technological and organisational infrastructure for systematic and controlled web harvesting and born-digital archiving
- ▶ to archive Slovak websites and born-digital content (electronic monographs and electronic serials) as an integral part of the Slovak Cultural Heritage

Organisational support:

- ▶ Deposit of Digital Resources (National ISSN Centre and Deposit of Digital Resources) – created in 2015
- ▶ external capacities: support – Service Level Agreement

DIGITAL RESOURCES - ORGANISATIONAL SUPPORT

2019 – the fourth year of sustainability of the project

Deposit of Digital Resources – staff:

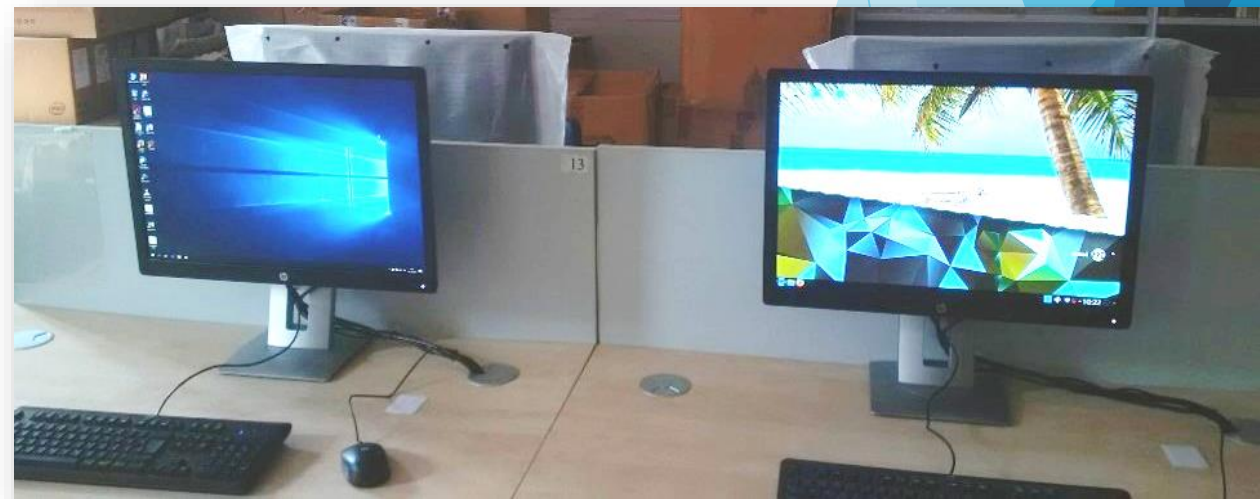
- ▶ Head of the Department (manager)
- ▶ Curator I - Administrator of the Information System for Digital Resources
- ▶ Curator II - Web Archive Curator
- ▶ Curator III – Born-Digital Archive Curator
- ▶ Manager and part-time person for Born-Digital Resources

DIGITAL RESOURCES - TECHNICAL INFRASTRUCTURE

- ▶ powerful HW infrastructure - Public and Internal Portal, 24 virtual servers
- ▶ system management is optimized for multiple harvesting processes (239 workers, each one has 15 Heritrix harvesters)
- ▶ there is an identical parallel testing environment, which enables test harvesting and problem analysing without interference of the production processes
- ▶ the system disposes with 800 TB storage
- ▶ application and system SW - open source technologies: Heritrix, DeDuplicator, OpenWayback, SOLR, JAVA, Invenio etc.
- ▶ support services: Backup, Monitoring, Communication, Antivirus...

ACCESS TO THE ARCHIVE

- ▶ according to the actual Copyright Act
- ▶ 3 types of access: public, local and forbidden
- ▶ default: without public access
- ▶ a limited number of archived websites and electronic publications is available publicly (Open Access, Creative Commons and agreements with provider/publisher)
- ▶ local access to the archived content in the ULIB research room

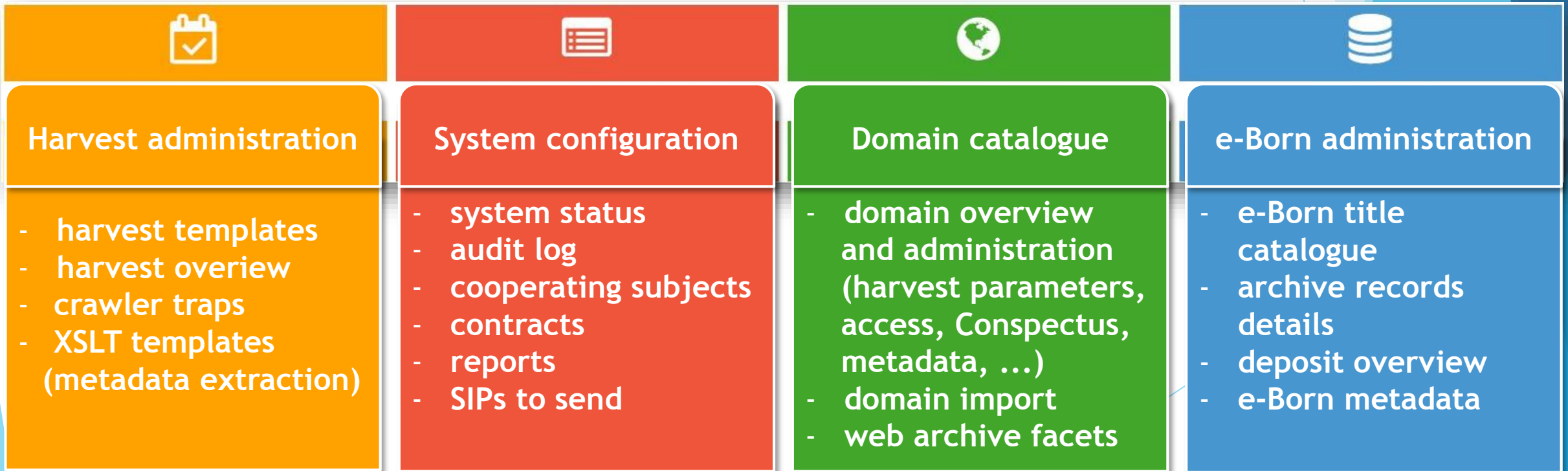


ARCHIVING OF ELECTRONIC PUBLICATIONS

- ▶ specific feature of our system – both websites and born-digital content (electronic serials and monographs with assigned ISSN/ISBN)
- ▶ 856 787 WARC files and 612 PDF files (14.6.2019)
- ▶ 144 serial titles, 59 monographs (14.6.2019)
- ▶ National ISSN Centre - cooperation from the beginning (2015)
- ▶ the Slovak National ISSN Database – the source of e-serials titles
- ▶ archiving is a preparation for the future Legal Deposit Act for born-digital content

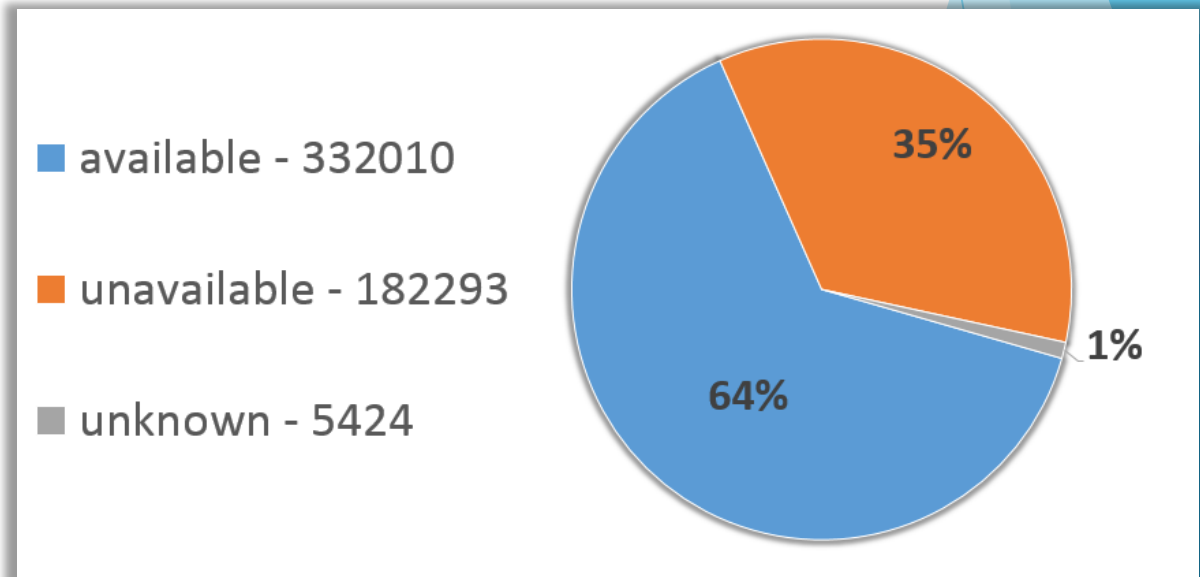
INTERNAL PORTAL - CURATOR INTERFACE

- ▶ 2x test virtual platform
- ▶ 1x production platform
- ▶ divided into 4 parts



WEB RESOURCES CATALOGUE

- ▶ nic.sk - current number of registered .sk domains 394 462 (14.06.2019)
- ▶ number of domains in the DIR catalogue: 519 727 (12.06.2019)
- ▶ other domains of slovacical character – content, author or language:
.eu (209), .com (109), .org (72) .info(35), .net(21), .cz (14),
other (23) - added into the catalogue manually
- ▶ Domain status in the catalogue:



WEB HARVEST POLICY

- ▶ respecting robots.txt
- ▶ agreements => respecting limitations
- ▶ harvest types: selective, thematic, full-domain (complex)
- ▶ formats: generally no audio, video files, streams, social networks

	2016	2017	2018	2019 done	2019 planned
Selective	7	11	14	5	9
Thematic	4	2	5	6	1
Full-domain	1	1	1	-	1
Σ	12	14	20	11	11

Table: Web harvest campaigns

HARVEST TYPES - 1: SELECTIVE HARVEST

- ▶ focus: Conspectus categories and contracted URLs
- ▶ contracted URLs - ignoring robots.txt, but respecting limitations in the agreement
- ▶ Conspectus - meta 072_7\$x
- ▶ agreement number - meta 542__n
- ▶ 2 GB, 3 days/ domain

Detail domény kniznica.kezmarok.sk/

Základné údaje Stav súhlasu Metadáta Konspekt História zberov História prieskumov Nastavenie zberu

Maximálny počet objektov	999999999
Maximálny počet bytov	2147483648
Maximálne trvanie zberu v sekundách	259200
Ignorovať robots.txt	Ignorovať
Konfiguračný súbor pre Heritrix	

Informácia o zmene

Dátum vytvorenia: 16.12.2016 13:21 | Vytvoril: phausleitner
Dátum poslednej zmeny: 8.12.2017 14:49 | Zmenil: phausleitner

Upraviť

Image: harvest settings for contracted domain

HARVEST TYPES - 2: THEMATIC HARVEST

- ▶ thematic (event) harvest
- ▶ 2 GB, 2 days/ domain
- ▶ elections (presidential, europarliamentary, regional), Olympic Games, IIHF Championship, ...
- ▶ respecting robots.txt

Source (accessible in ULIB):

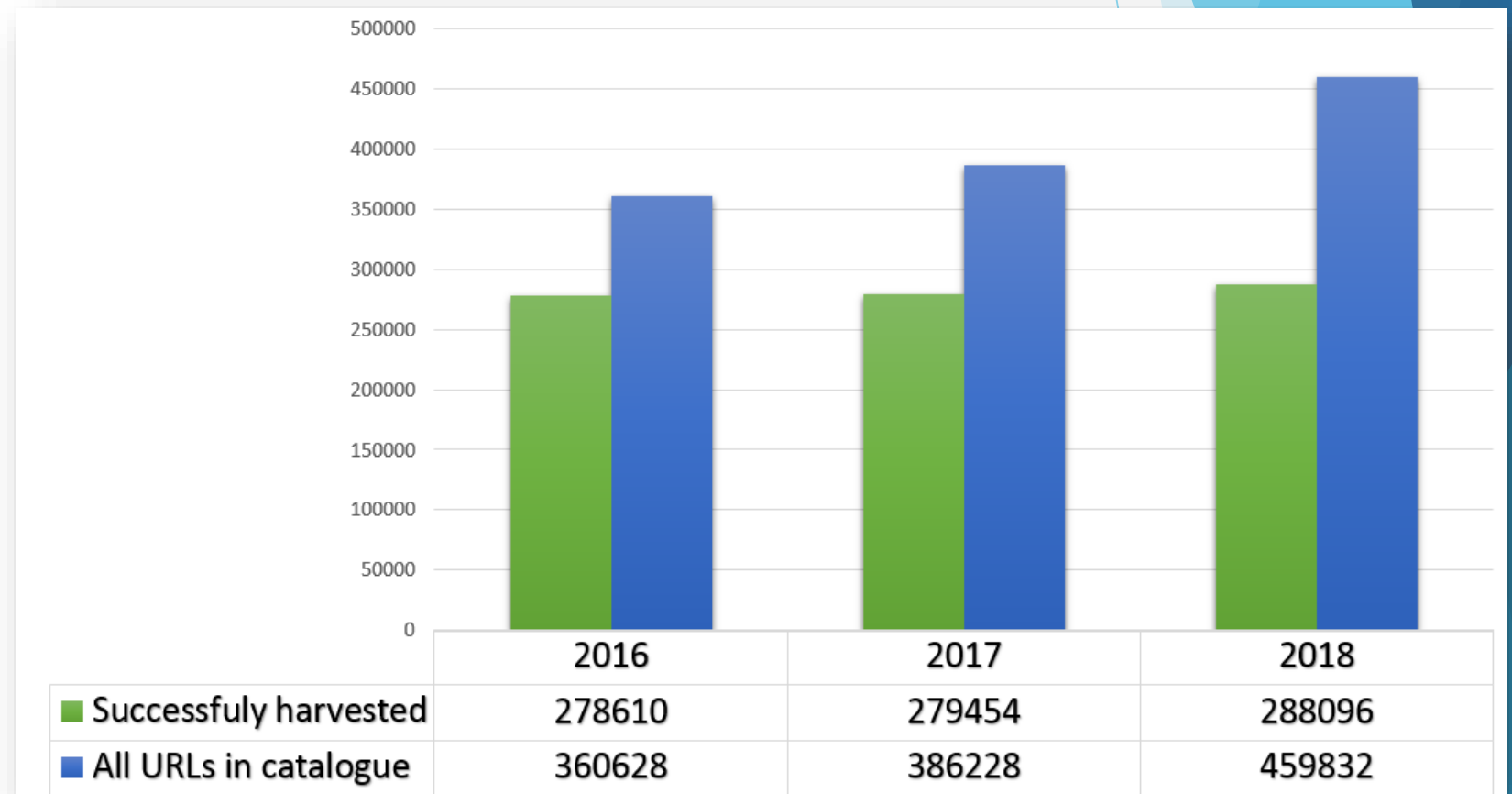
https://search.webdepozit.sk/webarchiv/kniznica/20190401151552im_/https://m.smedata.sk/api-media/media/image/sme/9/41/4109789/4109789_600x400.jpeg?rev=3

archived



HARVEST TYPES - 3: FULL-DOMAIN HARVEST

- ▶ harvest of all domains in the catalogue
- ▶ once a year
- ▶ parameters/ one seed:
 - ▶ max. size 400 MB
 - ▶ 10 000 objects
 - ▶ max. time 2 hrs.



PUBLIC PORTAL

- ▶ a lot of information for publishers including forms for publication submission/suggestion
- ▶ frontend for web archive and e-publications archive
- ▶ plenty of information about our activities
- ▶ made in Wordpress
- ▶ simple and advanced search

The screenshot shows the public portal interface. At the top left, there are accessibility options: A++ | A+ | A | Full text. The main header features the logo for 'digitálne pramene' (Digital Sources) on the left, a search bar with the text 'Vyhadavanie vo webovom archíve' (Searching in the web archive) in the center, and the logo for 'univerzitná knižnica v bratislave' (University Library in Bratislava) on the right. Below the search bar, there are radio buttons for 'www pramene' (selected) and 'elektronické publikácie' (electronic publications). A navigation menu below the header includes: SYSTEM DIR, ARCHIVES & CATALOGUES, WWW RESOURCES, E-PUBLICATIONS, DIR PROJECTS, CONTACTS, and SLOVENČINA. The main content area has a dark blue banner with the text 'We are working to have the Slovak web available for future generations' above a large image of a glowing blue map of Slovakia. To the right, there is a 'Webdepozit statistics' section with a donut chart. The chart shows three categories: 'Všetchné 766 TB' (All 766 TB) in green, 'Využitie 32 TB' (Usage 32 TB) in dark blue, and 'Celkovo 800 TB' (Total 800 TB) in purple. Below the chart are four empty square icons and a button labeled 'Archive profile'.

<https://webdepozit.sk>

SIMPLE SEARCH

- ▶ search in all metadata fields
- ▶ filtered by Conspectus categories
- ▶ links to live websites and archived versions

Screenshot of the search results page for the keyword "ulib.sk". The page title is "Výsledky vyhľadávania - web archiv (nájdenných 10 záznamov)". The search bar shows "ulib.sk" and buttons for "Hľadať" and "Zruš filter". The first result is for "www.ulib.sk / Univerzitná knižnica - Hlavná stránka". It includes a thumbnail of the website, the current version link, and a list of 29 archived versions with dates ranging from 2015 to 2018. A "Úplný záznam" link is provided. Below the result is a "Metadáta" section with a link to "Exportuj metadáta ako MARC, MARCXML".

Search of the keyword „ulib.sk“ – links to live site, OW versions, possible export of harvest metadata

Screenshot of the search results page for the keyword "Múzeum". The page title is "Výsledky vyhľadávania - web archiv (nájdenných 29 záznamov)". The search bar shows "Múzeum" and buttons for "Hľadať" and "Zruš filter". The results list includes:

- muzeum-malacky.blogspot.sk / Múzeum Michala Tíllnera**: Aktuálna verzia stránky, Archivované verzie stránky (2): 25.5.2018, 5.5.2018. Úplný záznam
- www.muzeumpezinok.sk/sk**: Aktuálna verzia stránky, Archivované verzie stránky (2): 24.2.2017, 7.10.2016. Úplný záznam
- www.zsmuzeum.sk/sk**: Aktuálna verzia stránky, Archivované verzie stránky (2): 24.2.2017, 7.10.2016. Úplný záznam
- www.muzeumbs.sk / Múzeum Banská Štiavnica**: Aktuálna verzia stránky, Archivované verzie stránky (6): 25.5.2018, 5.5.2018, Viac... Úplný záznam
- www.muzeumkn.sk / Podunajské múzeum v Komárne**

ADVANCED SEARCH

▶ Basic metadata search by following MARC fields:

- ▶ All metadata
- ▶ Title (24500\$a)
- ▶ Keyword (650#4\$a)
- ▶ Conspectus category (072#7\$x)
- ▶ URL of resource (85640\$u)
- ▶ Subtitle (24616\$i)
- ▶ Description (500##\$a)

▶ SOLR complex expression search using operators

- ▶ Logical operators:
 - ▶ and (AND)
 - ▶ or (OR)
 - ▶ not (NOT)

▶ filtering results by date of archiving

The screenshot displays a search interface with the following elements:

- Search Bar:** "Všetky kategórie" dropdown, "Hľadaný výraz" input field, "Hľadať" button, and "Zruš filter" button.
- Message:** "Výraz pre vyhľadávanie neobsahuje (po normalizácii) dostatočne dlhé slovo" (The search expression does not contain a sufficiently long word after normalization).
- Syntax:** "Syntax výrazu" with radio buttons for "jednoduchá" (selected) and "SOLR".
- MARC Search:** "Vyhľadať v MARC" dropdown menu with options: "Všade", "Hlavný názov (24500\$a)", "Kľúčové slovo (650#4\$a)", "Slovné vyjadrenie predmetovej kategórie (072#7\$x)", "URL prameňa (85640\$u)", "Variantný názov sídla (24616\$i)", and "Všeobecná poznámka (500##\$a)".
- Date Archiving:** "Dátum archivácie" with radio buttons for "Nezávisle od dátumu" (selected) and "od [] do []".
- Search Results:** "Výsledky vyhľadávania - web archív (nájdenných 2 763 záznamov)".
- Result Card:** Includes a thumbnail of a website, the URL "spis.eu.sk / Región Spiš - Spis.eu.sk - Úvod", a list of locations "Belá, Gelnica, Kežmarok, Levoča, Poprad, Spišská Nová Ves, Stará Ľubovňa, Svit, Vluchy, chata, fotogra...", a link to "Aktuálna verzia stránky", and "Archivované verzie stránky (1): 22.5.2018".
- Buttons:** "Zrušiť" and "Úplný záznam" buttons.

OPENWAYBACK

► Results in OpenWayback

► Displaying of the archived website

Screenshot of the OpenWayback search results page for the URL <https://www.ulib.sk/sk/>. The page shows a search bar with the URL, a search button labeled "Hľadať", and a bar chart indicating the number of captures over time. The bar chart shows a significant increase in captures starting in 2018, peaking in 2019. Below the chart is a calendar for the year 2019, with the date 20th of April highlighted in blue, indicating the selected snapshot. The page also includes the OpenWayback logo and the text "digitálne pramene".

Screenshot of the archived website <https://www.ulib.sk/sk/>. The page displays the website's header, navigation menu, and main content area. The header includes the OpenWayback logo and the text "digitálne pramene". The navigation menu includes links for "Katalógu UKB", "Metalibe", "na stránkach", and "v google". The main content area features a banner for "Lesy a roviny" by Michal Ďurovka, a section for "Profil slovenskej kultúry" with a list of categories, and a "Novinky" (News) section with several articles. The footer includes social media links, a virtual visit button, and copyright information.

<https://search.webdepozit.sk>

E-BORN SEARCH

- ▶ full-text search
- ▶ results of e-Born archive search with title details
- ▶ title „Dejiny“



NASPÁŤ NA WEBDEPOZIT.SK

Jednoduché vyhľadávanie | Pomoc

História a pomocné v

▼

dejiny

Hľadaj

Zruš filter

Syntax výrazu

jednoduchá

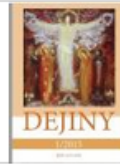
SOLR

Vyhľadať v MARC

Názov (24500\$a)

Výsledky vyhľadávania - eBorn archív (nájdeneých 6 záznamov)

Dejiny.; Rok 2015, Roč. 10, č. 1



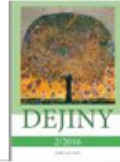
Dejiny.; Rok 2015, Roč. 10, č. 1

UNIVERSUM-EU, 2015 (číslo elektronického seriálu)

SICI: 1337-0707(2015)10:1<ID:DDP-EB000000984320652>3.0.CO;2-T

Súbory: Dejiny_(2015)10_1.pdf (9.1 MB), Dejiny_2015_1_Obalka.pdf (917 KB)

Úplný záznam



Dejiny.; Rok 2016, Roč. 11, č. 2

UNIVERSUM-EU, 2016 (číslo elektronického seriálu)

SICI: 1337-0707(2016)11:2<ID:DDP-EB000000984327389>3.0.CO;2-0

Súbory: Dejiny_(2016)11_2.pdf (13 MB), Dejiny_2016_2_obalka.pdf (945 KB)

Úplný záznam

Metadáta čísla

Titul

Ďalšie čísla

Vyd. údaje:

- miesto vydania: Prešov :

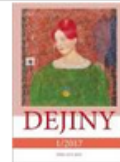
- vydavateľ: UNIVERSUM-EU s.r.o.;

Periodicita: 2x ročne

ISSN: 1337-0707

URL: <http://dejiny.unipo.sk/>

Exportuj metadáta titulu ako MARC, MARCXML



Dejiny.; Rok 2017, Roč. 12, č. 1

UNIVERSUM-EU, 2017 (číslo elektronického seriálu)

SICI: 1337-0707(2017)12:1<ID:DDP-EB000000984327766>3.0.CO;2-Z

Súbory: Dejiny_1_2017.pdf (12 MB), Dejiny_2017_1_obalka.pdf (934 KB)

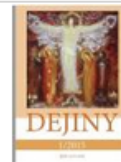
Úplný záznam

Metadáta čísla

Titul

Ďalšie čísla

Rok vydania: 2015

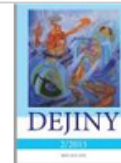


Dejiny.; Rok 2015, Roč. 10, č. 1

UNIVERSUM-EU, 2015 (číslo elektronického seriálu)

SICI: 1337-0707(2015)10:1<ID:DDP-EB000000984320652>3.0.CO;2-T

Súbory: Dejiny_(2015)10_1.pdf (9.1 MB), Dejiny_2015_1_Obalka.pdf (917 KB)

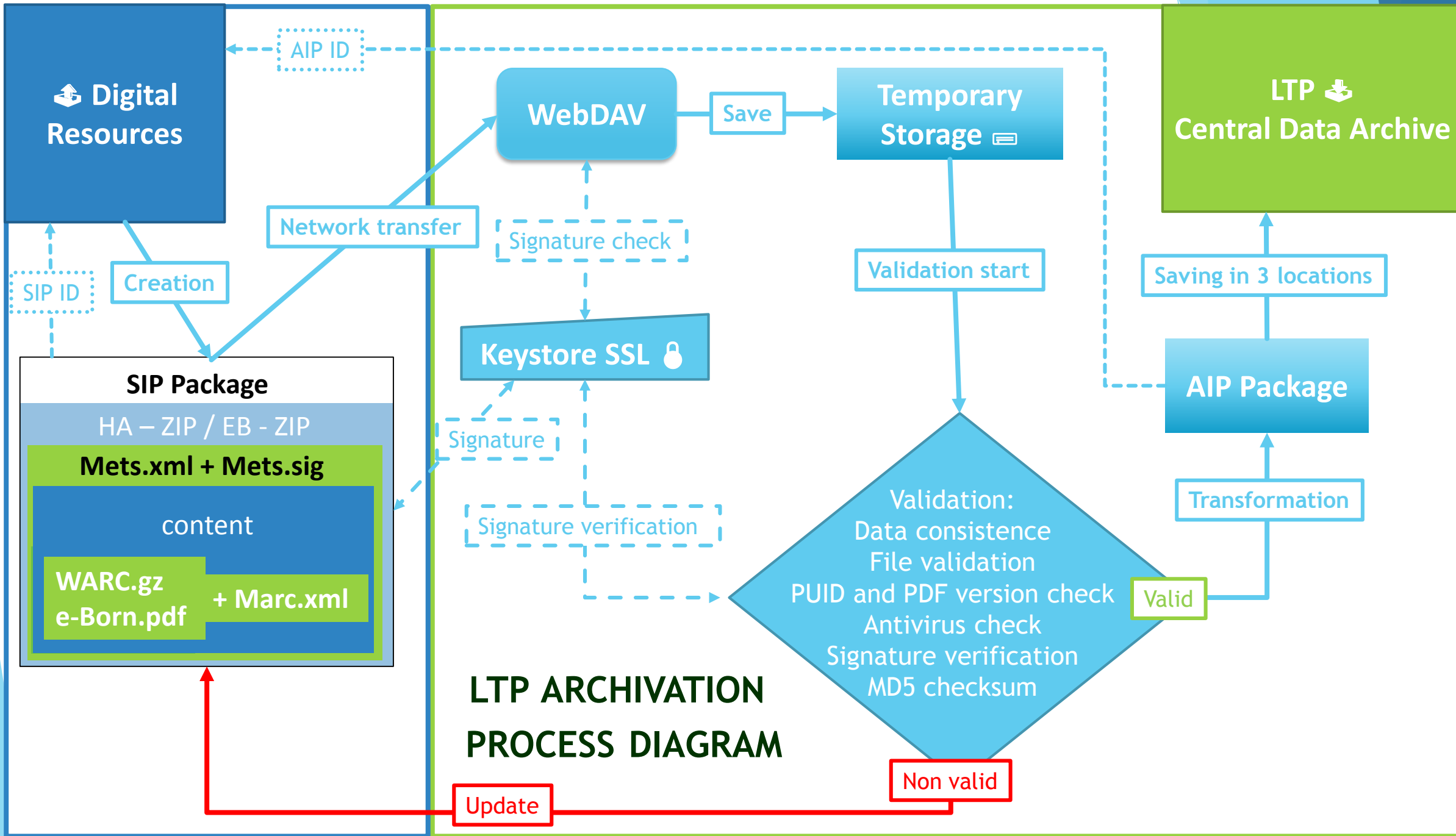


Dejiny.; Rok 2015, Roč. 10, č. 2

UNIVERSUM-EU, 2015 (číslo elektronického seriálu)

SICI: 1337-0707(2015)10:2<ID:DDP-EB000000984327238>3.0.CO;2-F

Súbory: Dejiny_(2015)10_2.pdf (13 MB), Dejiny_2015_2_Obalka.pdf (810 KB)



THANK YOU FOR YOUR ATTENTION

PhDr. Jana Matúšková & Ing. Peter Hausleitner
University Library in Bratislava, Slovakia
Deposit of Digital Resources (ddp@ulib.sk)

