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

digitálne pramene

ÚRAD VLÁDY SLOVENSKEJ REPUBLIKY

MINISTERSTVO KULTÚRY SLOVENSKEJ REPUBLIKY

EURÓPSKA ÚNIA

OPERAČNÝ PROGRAM INFORMATIZÁCIA SPOLOČNOSTI

TVORÍME VEDOMOSTNÚ SPOLOČNOSŤ

univerzitná knižnica v bratislave

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

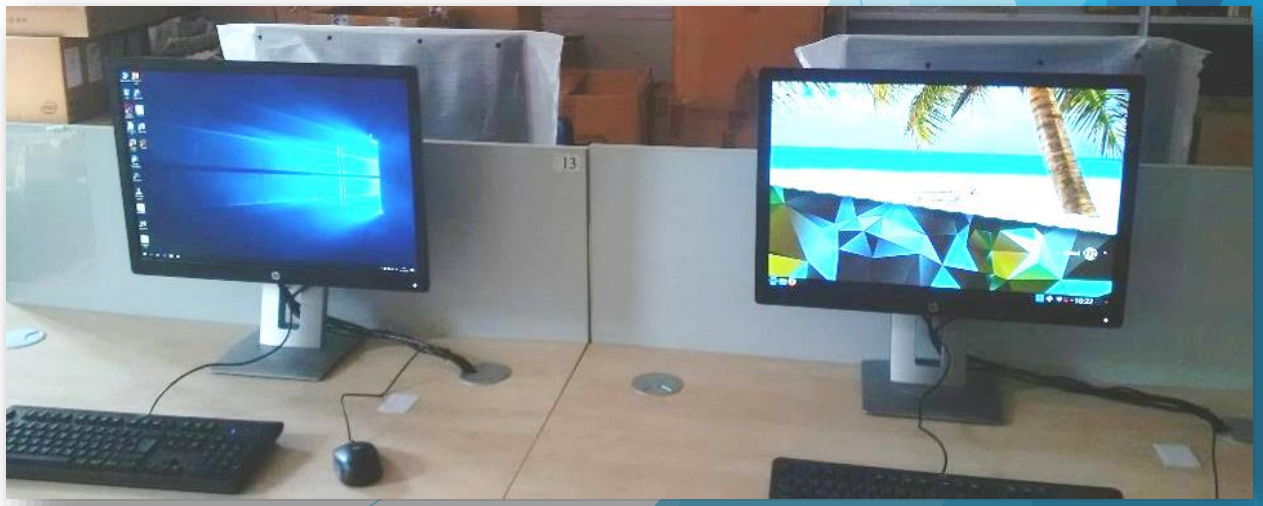
DIGITAL RESOURCES – TECHNICAL INFRASTRUCTURE

- 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

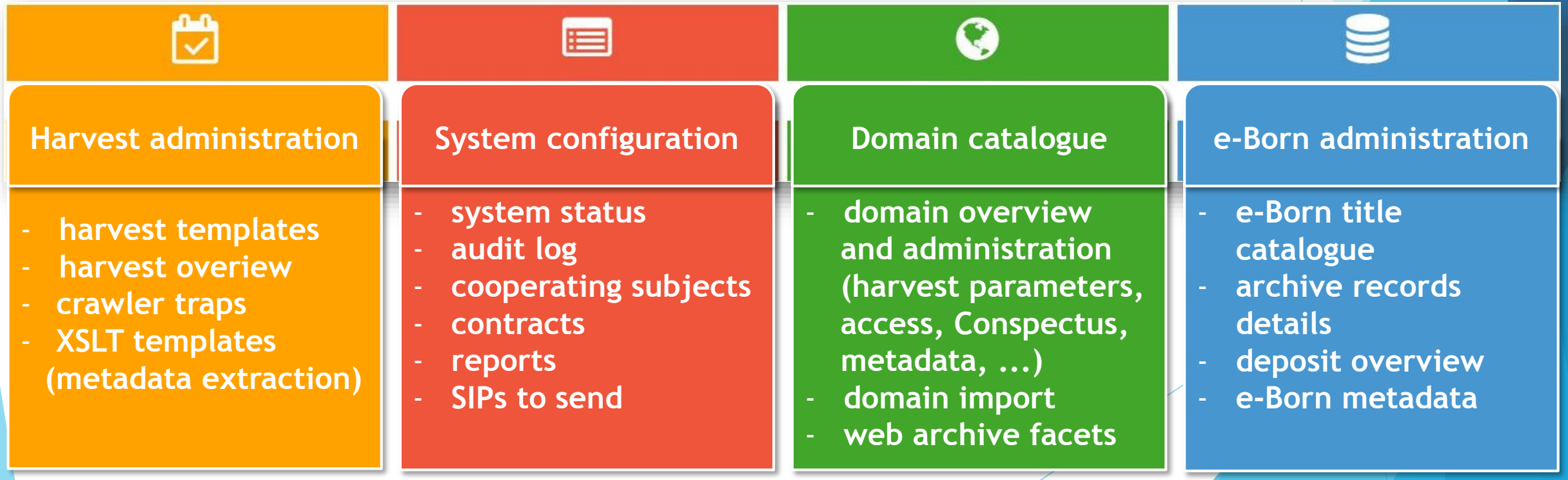


ARCHIVING OF ELECTRONIC PUBLICATIONS

- 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

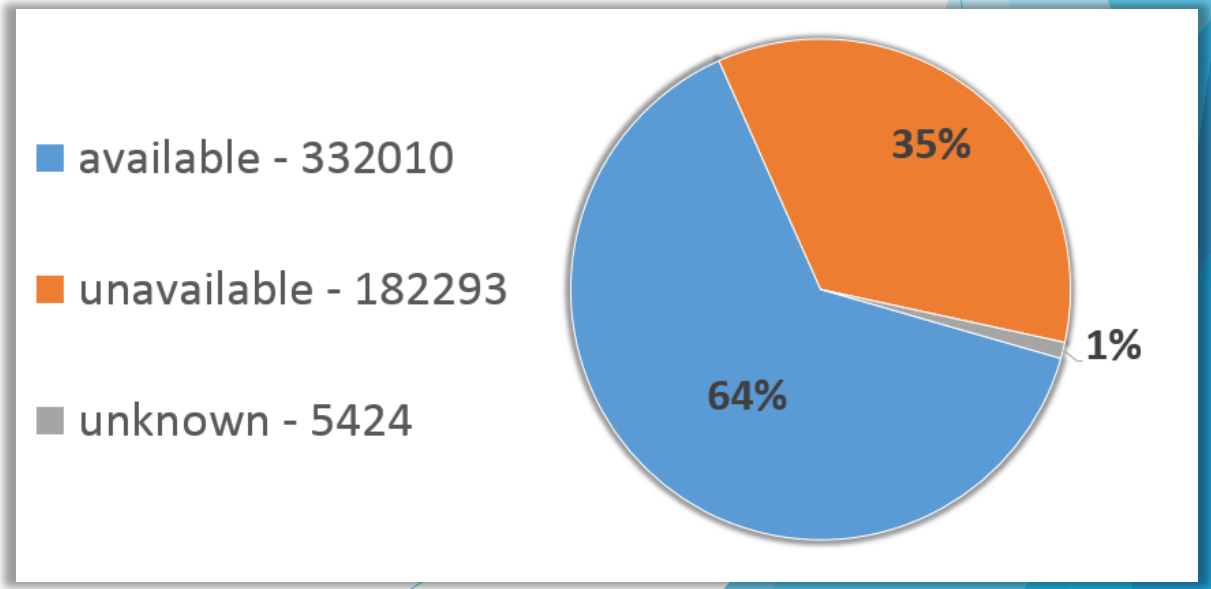


WEB RESOURCES CATALOGUE

- 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

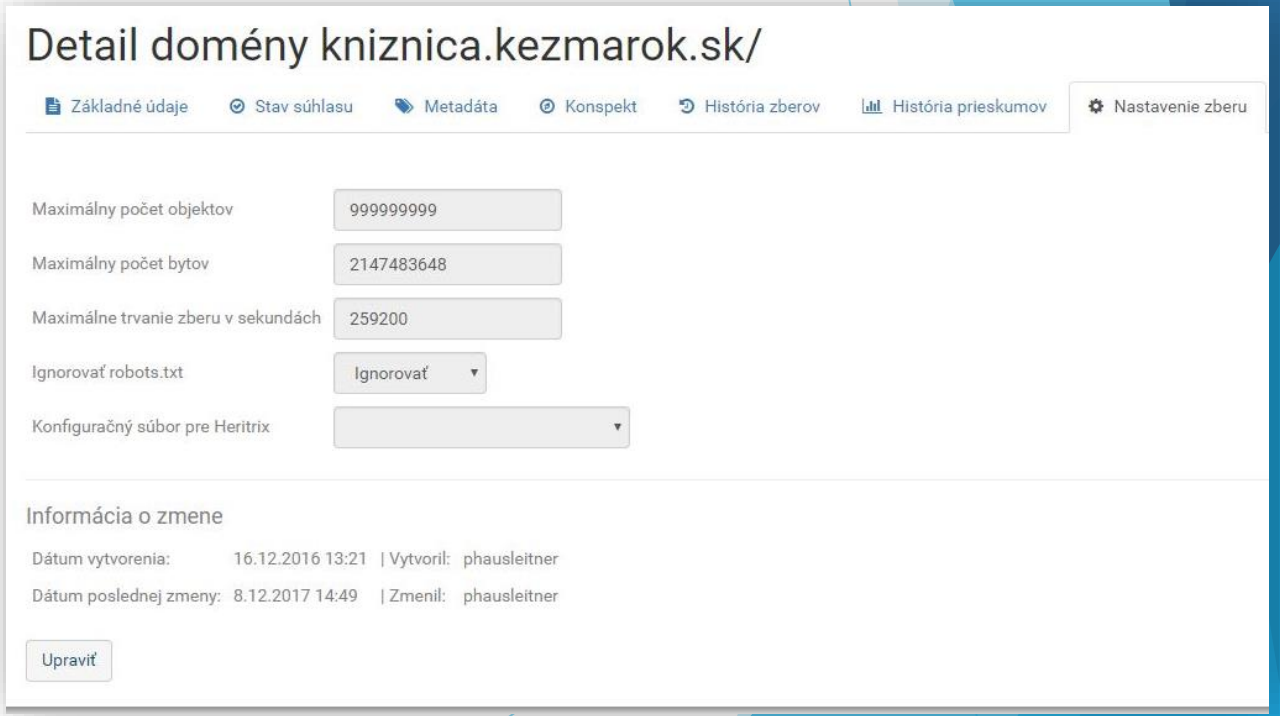


Image: harvest settings for contracted domain

HARVEST TYPES – 2: THEMATIC HARVEST

- thematic (event) harvest
- 2 GB, 2 days/ domain
- elections (presidential, europarliamentary, regional),

Olympic Games, IIHF Championship, ...

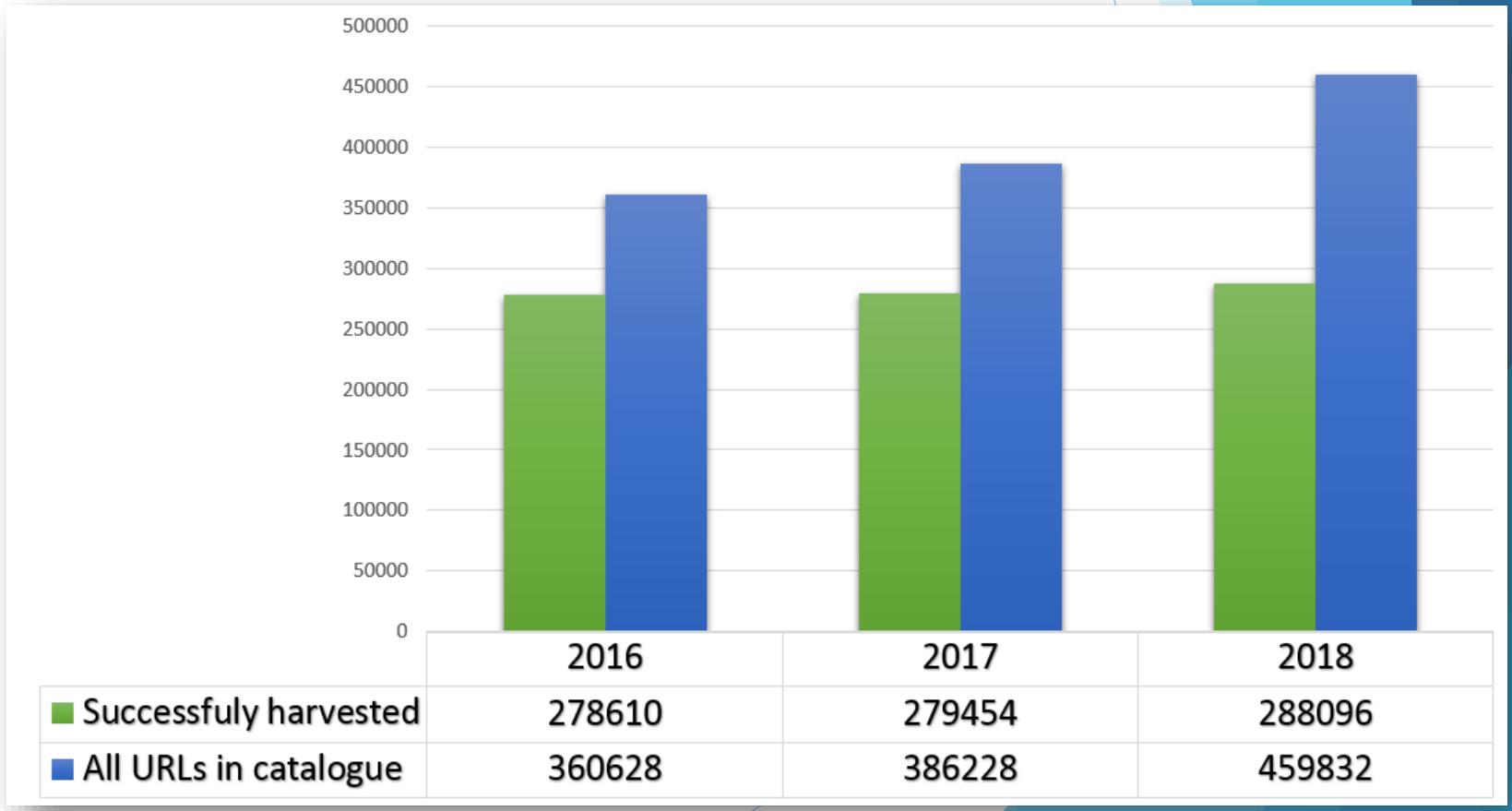
- respecting robots.txt

archived

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

HARVEST TYPES – 3: FULL-DOMAIN HARVEST

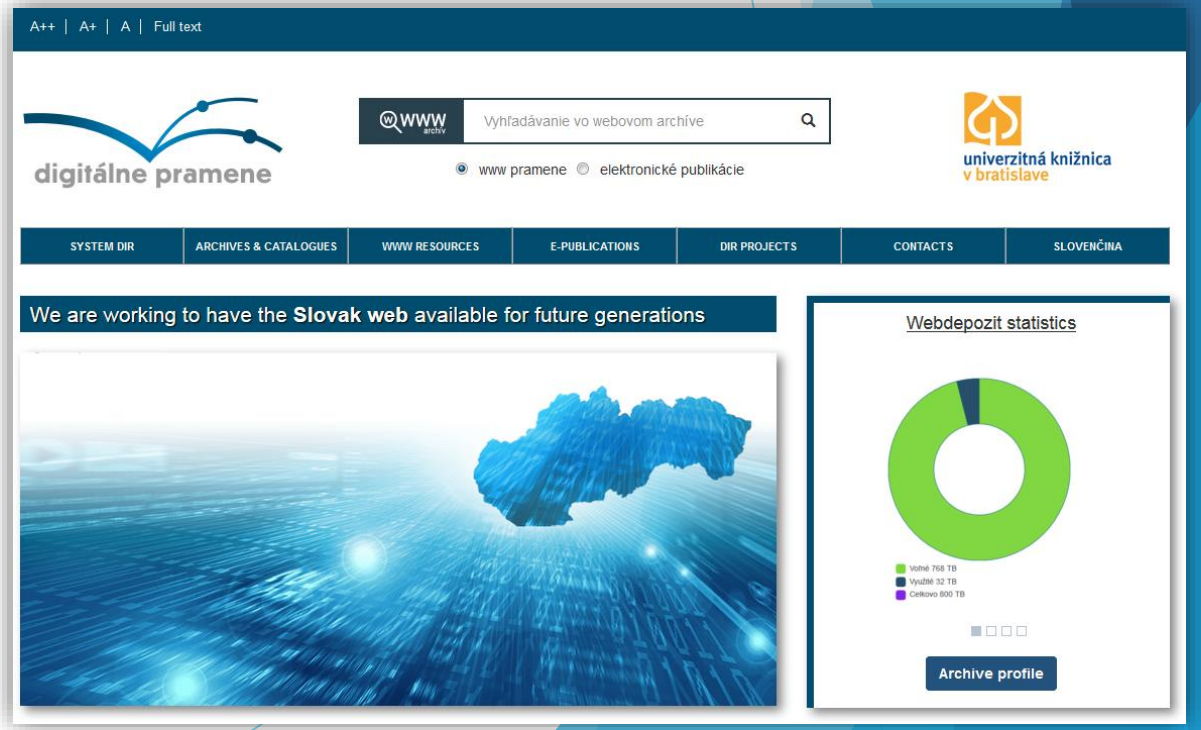
- harvest of all domains in the catalogue
- once a year
- parameters/ one seed:
  - max. size 400 MB
  - 10 000 objects
  - max. time 2 hrs.



	2016	2017	2018
Successfuly harvested	278610	279454	288096
All URLs in catalogue	360628	386228	459832

PUBLIC PORTAL

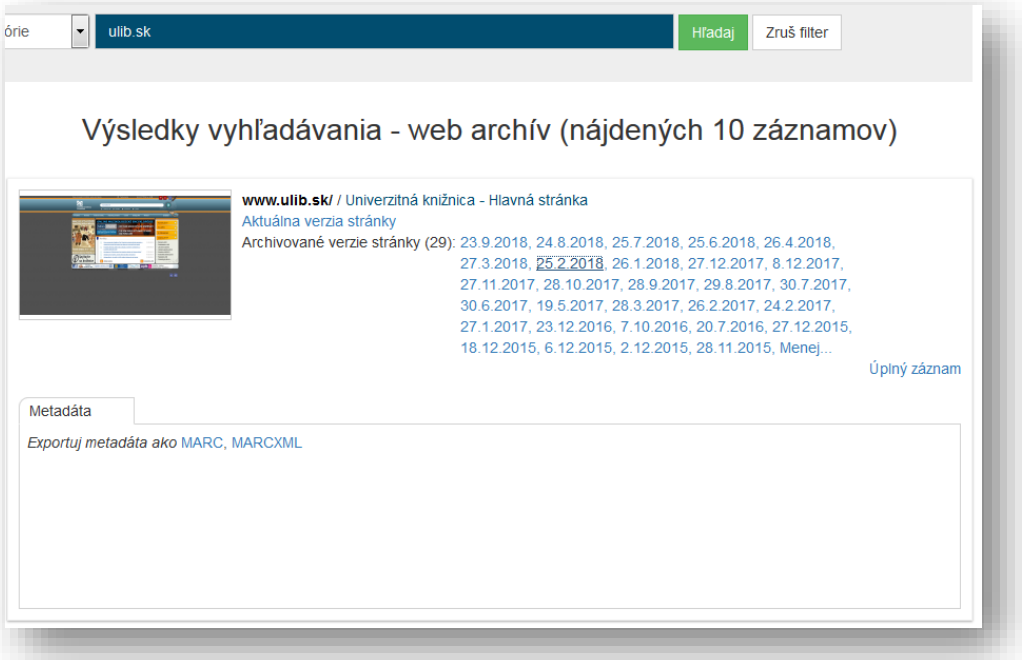
- a lot of information for publishers including forms for publication submission/suggestion
- frontend for web archive and e-publications archive
- plenty of information about our activities
- made in Wordpress
- simple and advanced search

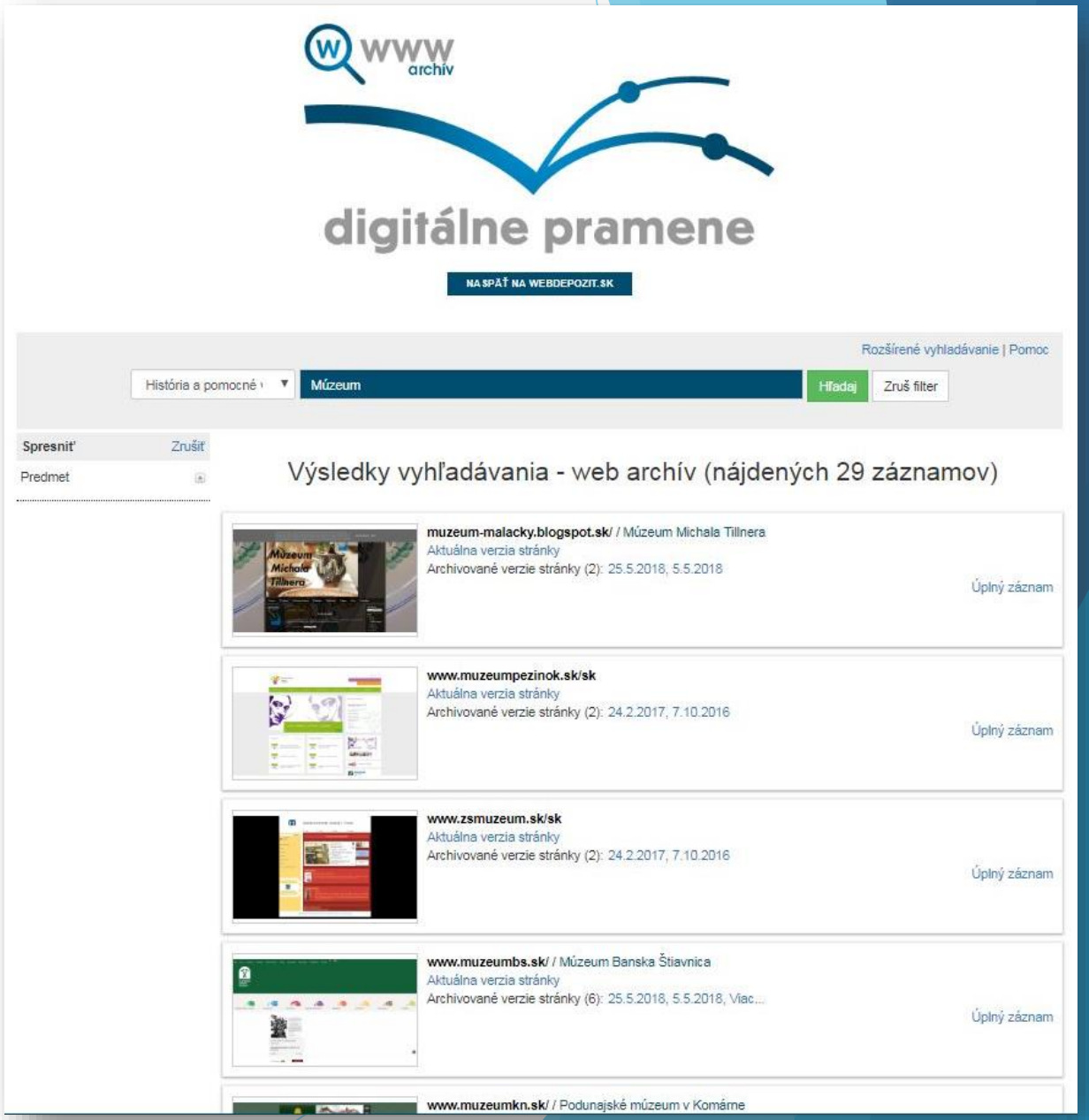


https://webdepozit.sk

SIMPLE SEARCH

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

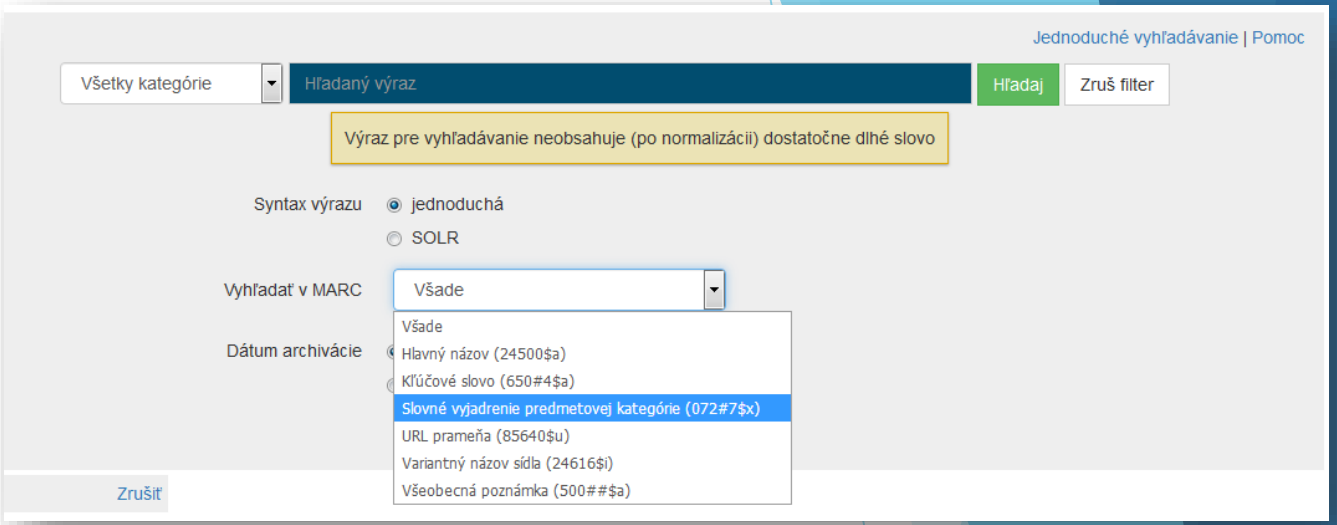


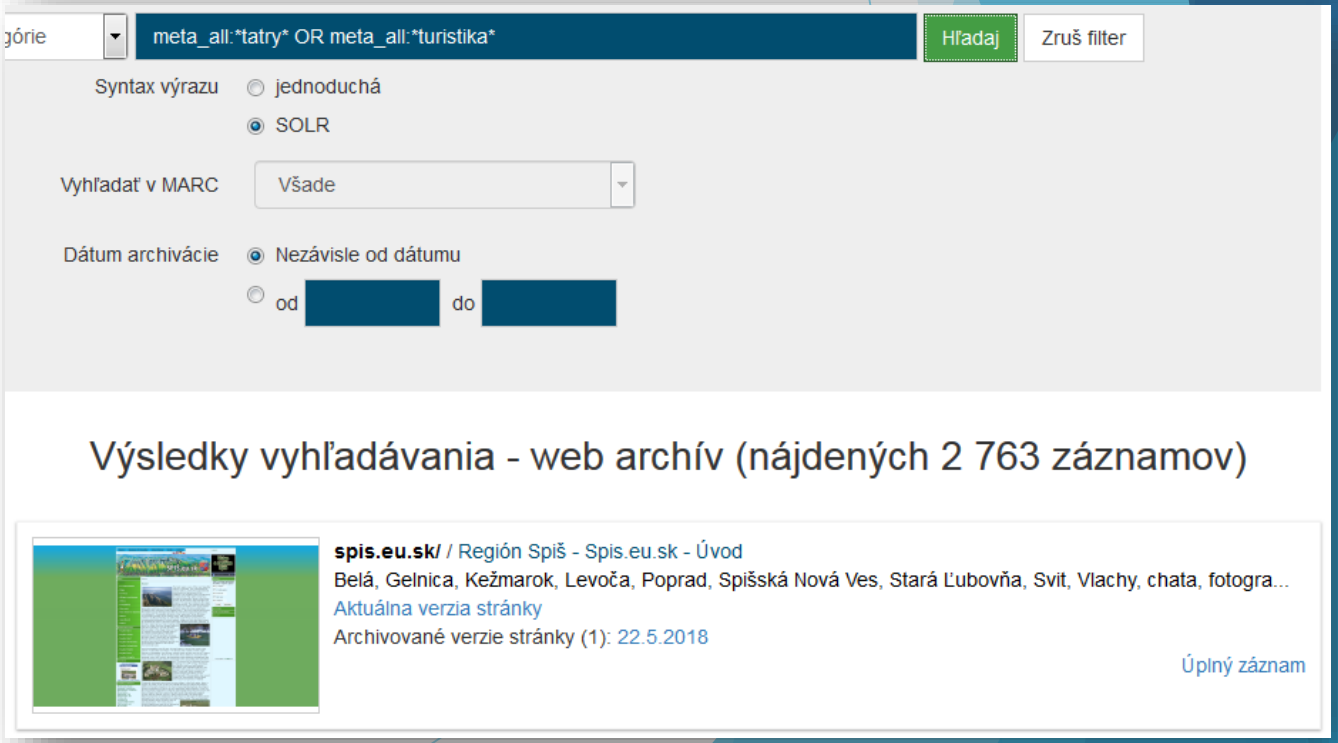


Search of the keyword „ulib.sk“ – links to live site, OW versions, possible export of harvest metadata

ADVANCED SEARCH

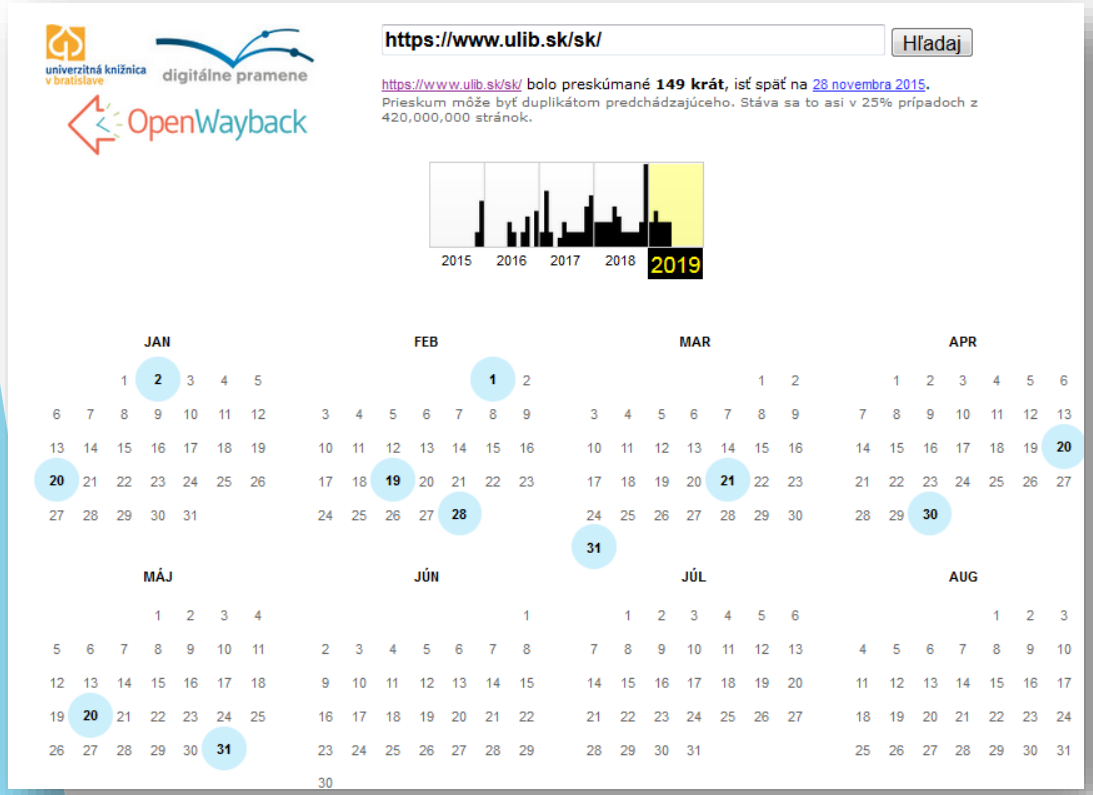
- Basic metadata search by following MARC fields:
  - All metadata
  - Title (24500$a)
  - Keyword (650#4$a)
  - Conspectus category (072#7$x)
  - URL of resource (85640$u)
  - Subtitle (24616$i)
  - Description (500##$a)
- SOLR complex expression search using operators
  - Logical operators:
    - and (AND)
    - or (OR)
    - not (NOT)
- filtering results by date of archiving





OPENWAYBACK

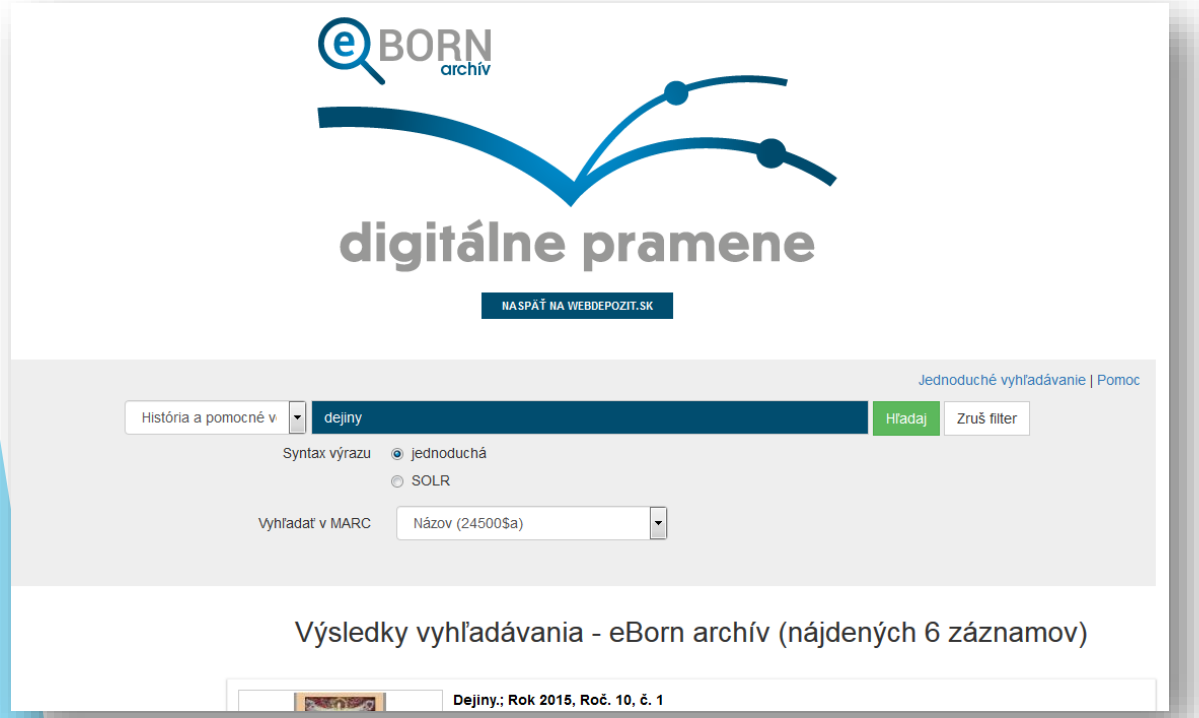
- Results in OpenWayback
- Displaying of the archived website

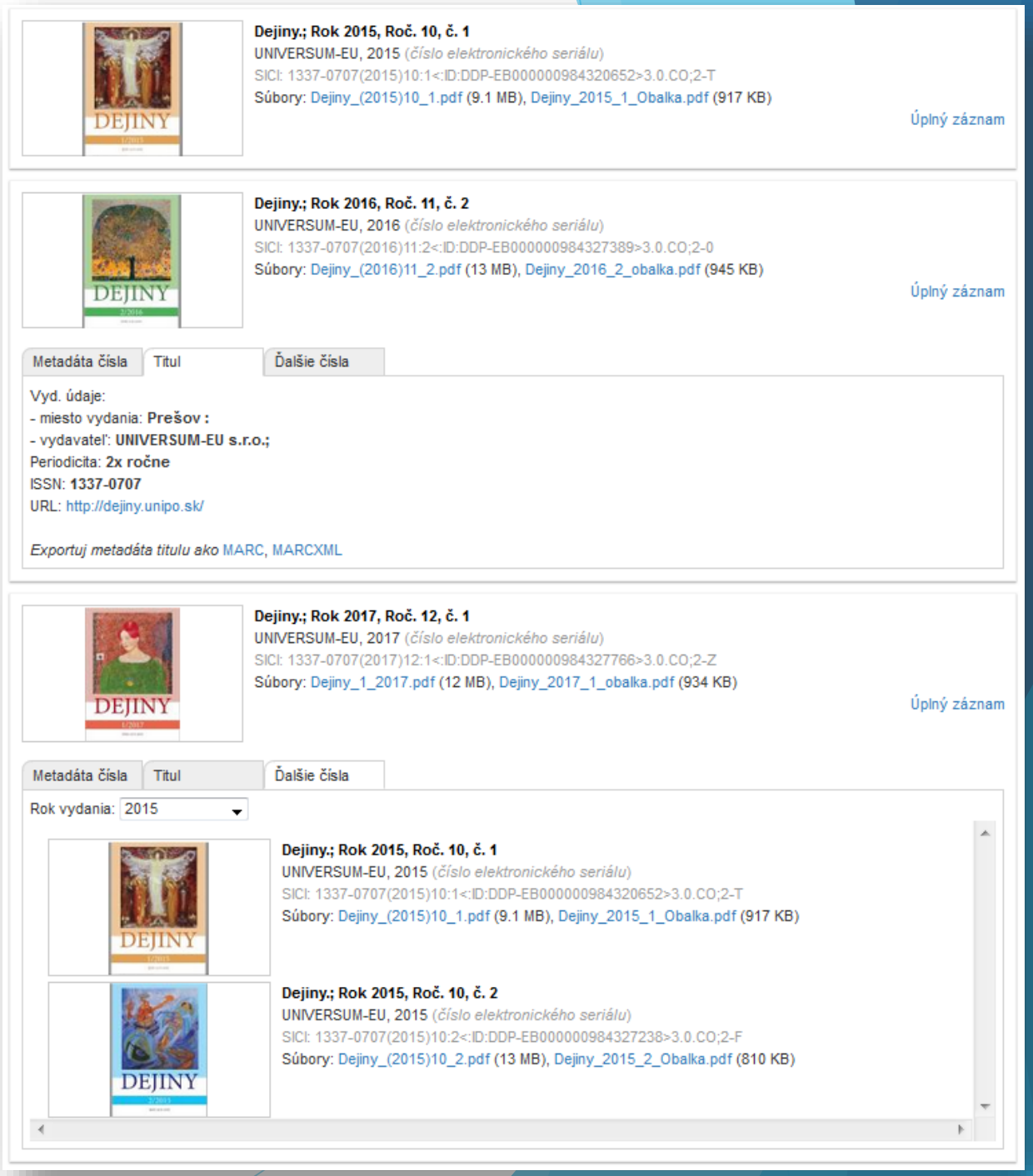


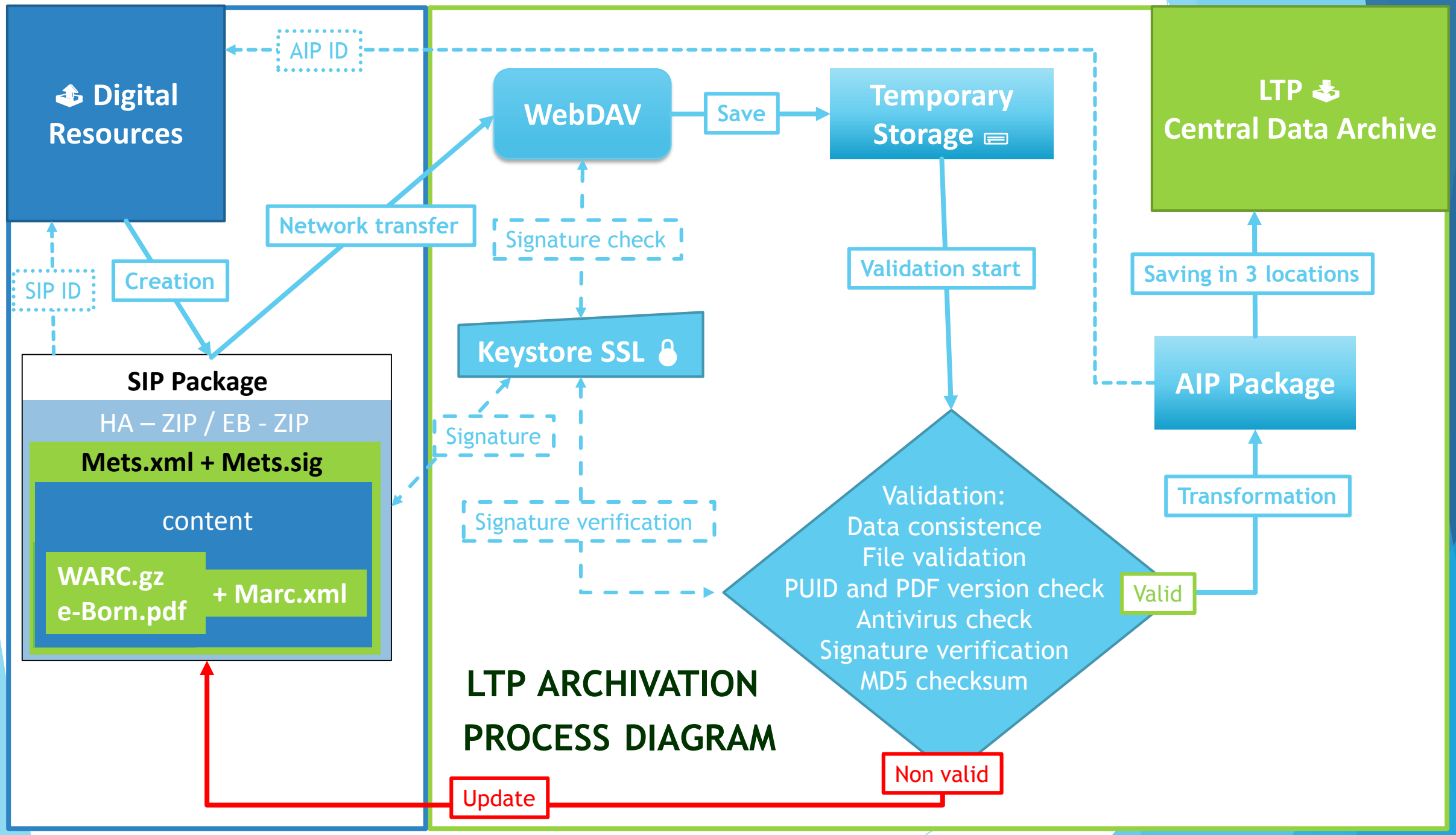
https://search.webdepozit.sk

E-BORN SEARCH

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






Digital Resources
AIP ID
WebDAV
Save
Temporary Storage
LTP
Central Data Archive
Network transfer
Signature check
Validation start
Saving in 3 locations
SIP ID
Creation
Keystore SSL
AIP Package
SIP Package
HA – ZIP / EB - ZIP
Mets.xml + Mets.sig
content
WARC.gz
e-Born.pdf
+ Marc.xml
Signature
Signature verification
Validation:
Data consistence
File validation
PUID and PDF version check
Antivirus check
Signature verification
MD5 checksum
Valid
Transformation
LTP ARCHIVATION
PROCESS DIAGRAM
Non valid
Update

THANK YOU FOR YOUR ATTENTION

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

digitálne pramene

ÚRAD VLÁDY SLOVENSKEJ REPUBLIKY

MINISTERSTVO KULTÚRY SLOVENSKEJ REPUBLIKY

EURÓPSKA ÚNIA

OPERAČNÝ PROGRAM INFORMATIZÁCIA SPOLOČNOSTI

TVORÍME VEDOMOSTNÚ SPOLOČNOSŤ

univerzitná knižnica v bratislave