DIGITAL RESOURCES – THE NATIONAL PROJECT OF WEB HARVESTING AND WEB ARCHIVING IN SLOVAKIA

Jana Matúšková, Peter Hausleitner, Andrej Bizík
University Library in Bratislava
IIPC WAC 2018, Wellington, New Zealand

digitálne pramene

ÚRAD VLÁDY SLOVENSKEJ REPUBLIKY

MINISTERSTVO KULTÚRY SLOVENSKEJ REPUBLIKY

EURÓPSKA ÚNIA

OPERAČNÝ PROGRAM INFORMATIZÁCIA SPOLOČNOSTI

TVORÍME VEDOMOSTNÚ SPOLOČNOSŤ

univerzitná knižnica v bratislave

UNIVERSITY LIBRARY IN BRATISLAVA

Michalská street

Franz Liszt´s Pavilion and Garden

Ventúrska st. – courtyard with glass ceiling

NATIONAL PROJECT DIGITAL RESOURCES

- web archiving in Slovakia - first attempts 2005/2006 - ULIB participated in project Web Cultural Heritage
- there were no necessary legislative, organisational and technical conditions for systematic harvesting and archiving of web pages and original electronic resources in Slovakia
- Operational Programme Informatization of the Society – 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:

- creation of the technological and organisational infrastructure for systematic archiving of web and e-Born (Born Digital) resources
- systematic collecting and long-time archiving of 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

2018 – the third year of sustainability of the Digital Resources Project

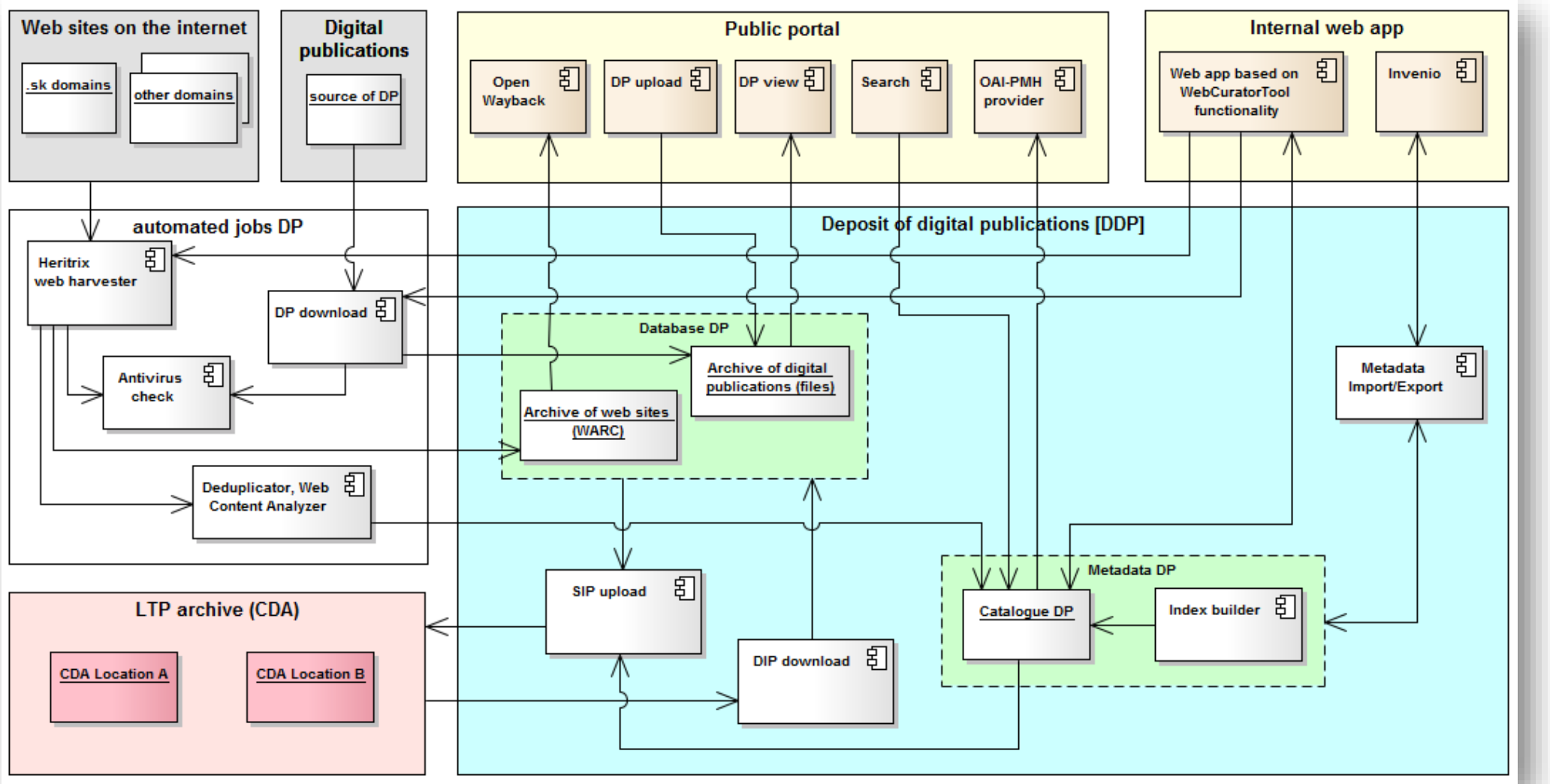
Deposit of Digital Resources – staff:

- 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 (Head of the Department National ISSN Centre and Deposit of Digital Resources)

DIGITAL RESOURCES – TECHNICAL INFRASTRUCTURE

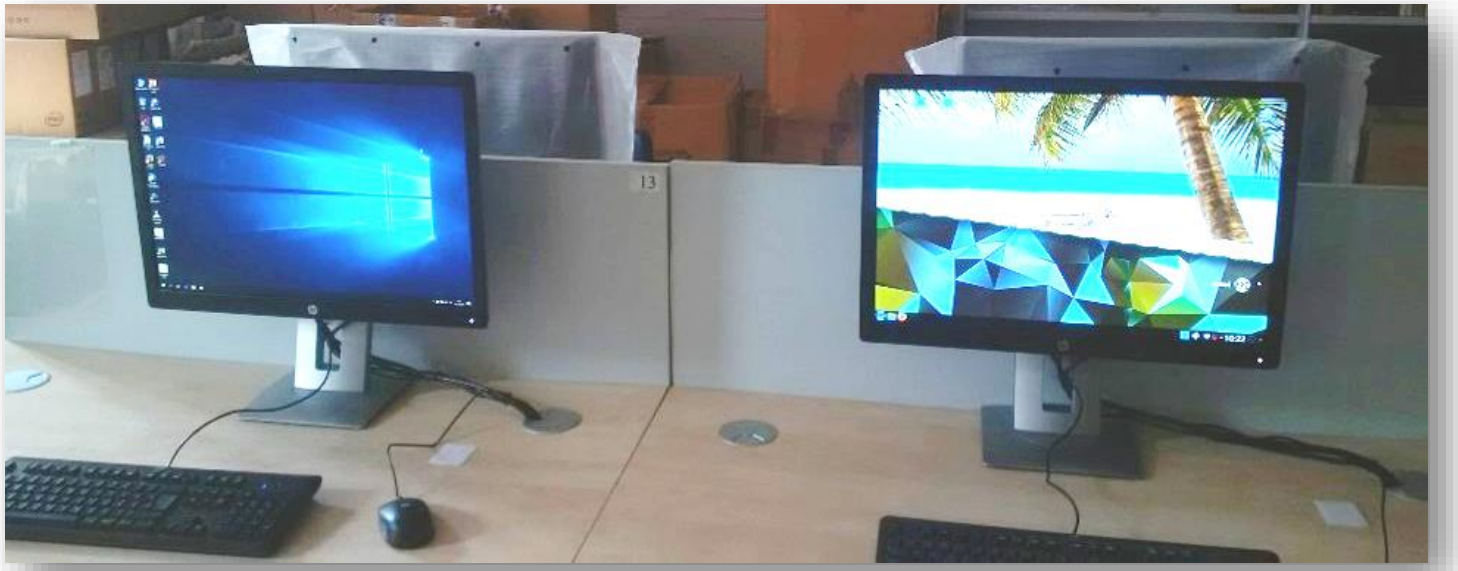
- powerful HW infrastructure - CISCO Server Cluster – Public and Internal Portal, 24 virtual servers
- system management is optimized for multiple harvesting processes (239 workers, each one has 15 Heritrix harvesters)
- there is an identical parallel testing environment, which enables test harvesting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

DIGITAL RESOURCES – ARCHITECTURE



ACCESS TO THE ARCHIVE

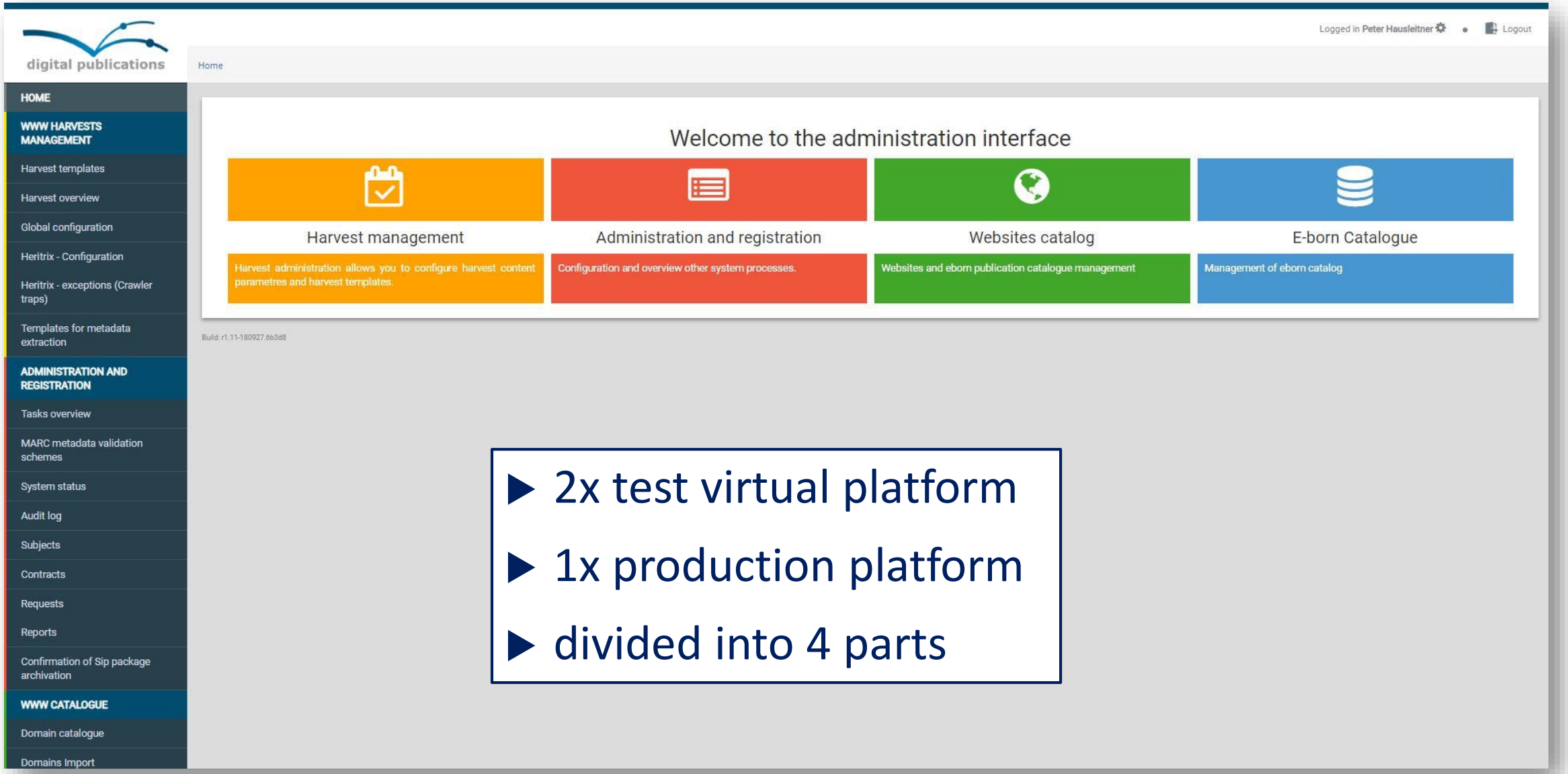
- according to the actual Copyright Act
- default: without public access
- Open Access, Creative Commons – no agreement necessary
- in other cases it is necessary to sign an agreement with provider/publisher
- 3 types of access: public, local and forbidden
- local access to the archived content in the ULIB research room (26 PCs)



ARCHIVING OF ELECTRONIC PUBLICATIONS

- specific feature of our system – both websites and Born Digital content (electronic serials and monographs)
- electronic publications are an endangered part of the national documentary heritage
- ISSN/ISBN is a precondition for saving to the archive
- National ISSN Centre - cooperation from the very 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

INTRA PORTAL - CURATOR INTERFACE



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

INTRA PORTAL - CURATOR INTERFACE

Harvest administration

- harvest templates
- harvest overiew
- crawler traps
- XSLT templates (metadata extraction)

System configuration

- system status
- audit log
- cooperating subjects
- contracts
- reports
- SIPs to send

Domain catalogue

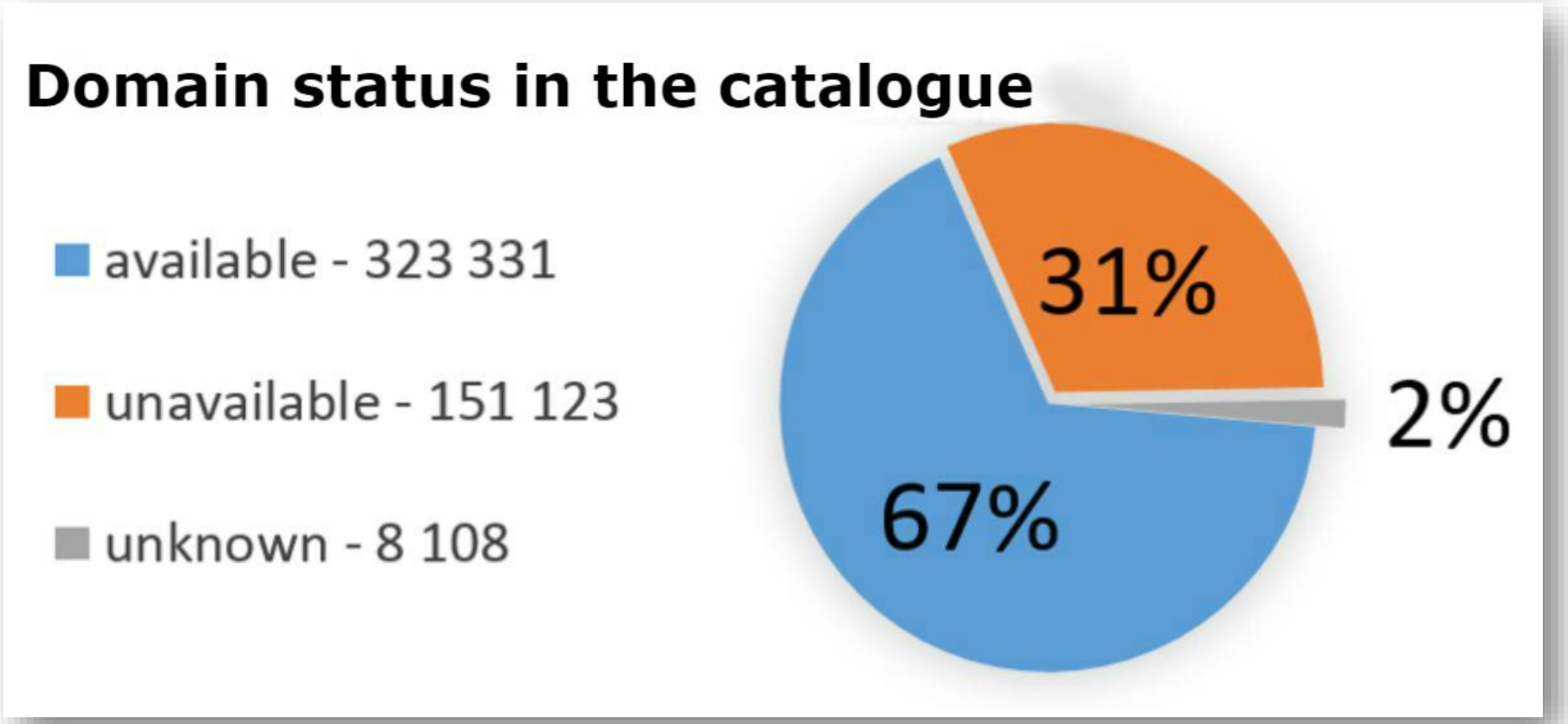
- domain overview and administration (harvest parameters, access, Conspectus, metadata, ...)
- domain import
- web archive facets

e-Born administration

- e-Born title catalogue
- archive records details
- deposit overview
- e-Born metadata

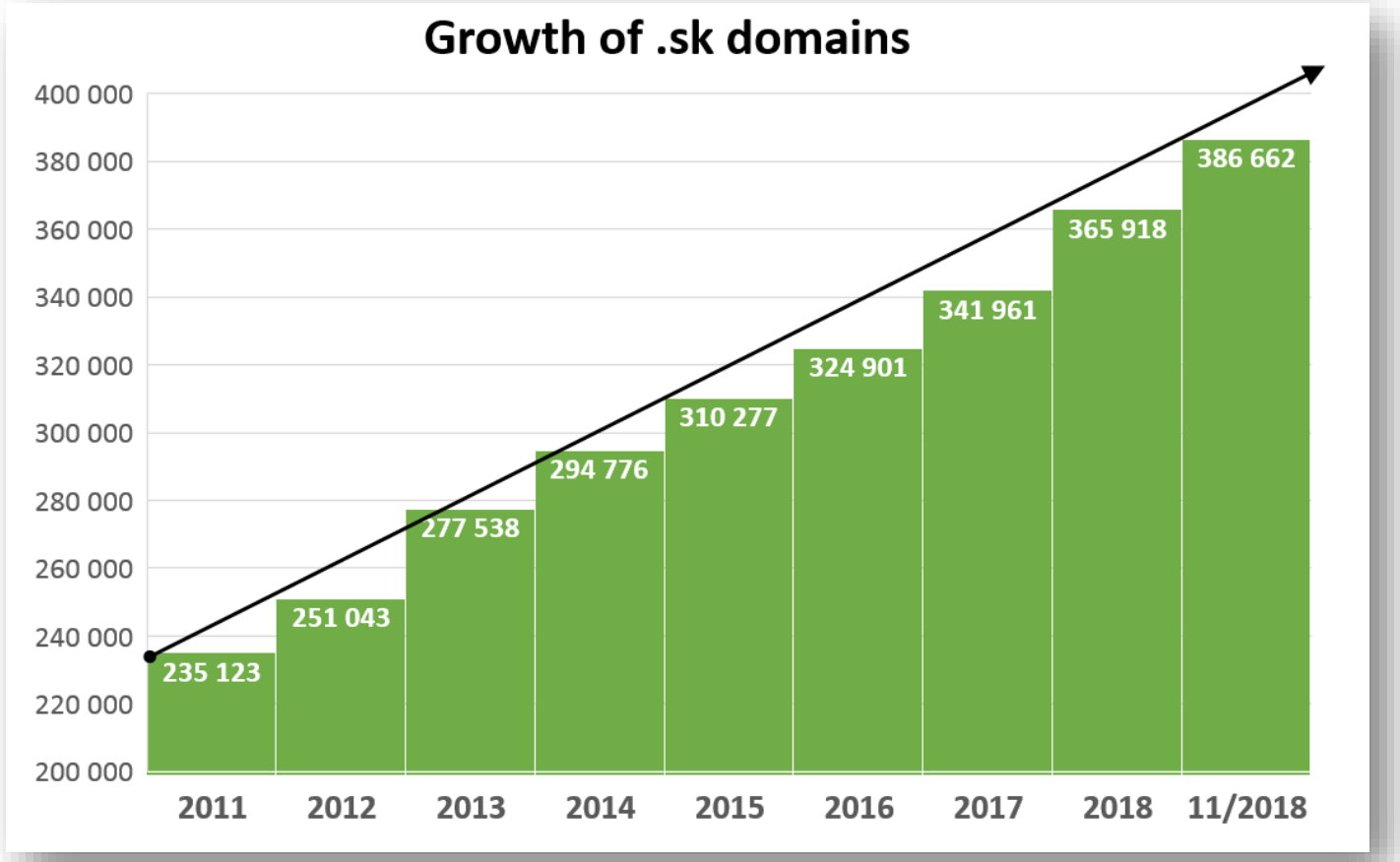
WEB RESOURCES CATALOGUE

number of domains in the catalogue: 482 562 (16.10.2018)



WEB RESOURCES

- nic.sk - current number of registered .sk domains
- other domains of slovacical character – content, author or language:
  - .com (67),
  - .eu (60),
  - .org (57)
  - .info(11),
  - .cz (10),
  - .net(7),
  - other (16)...
  - are added into the catalogue manually



WEB HARVEST POLICY

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

	2015	2016	2017	2018
Selective	1	7	11	12
Thematic	1	4	2	5
Full-domain	1	1	1	1
Σ	**3**	**12**	**14**	**18**

Table: Web harvest campaigns until October 2018

HARVEST TYPES - 1: SELECTIVE HARVEST

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

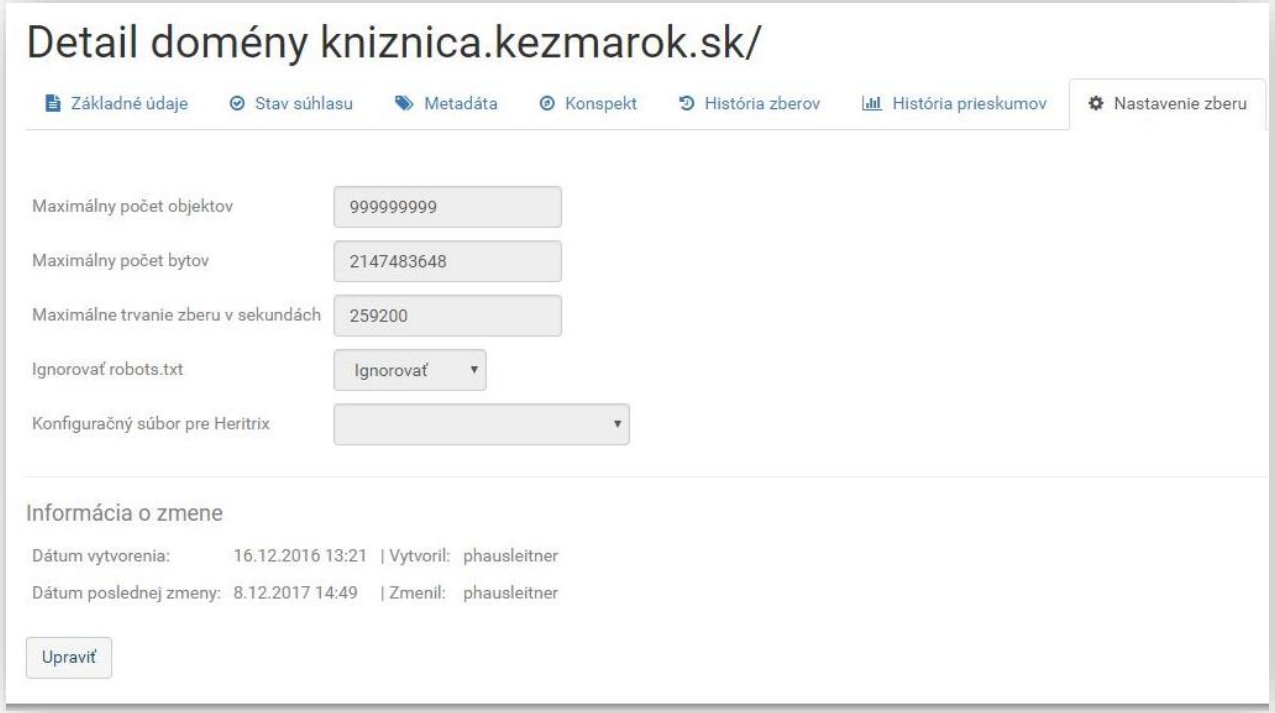


Image: harvest settings for contracted domain

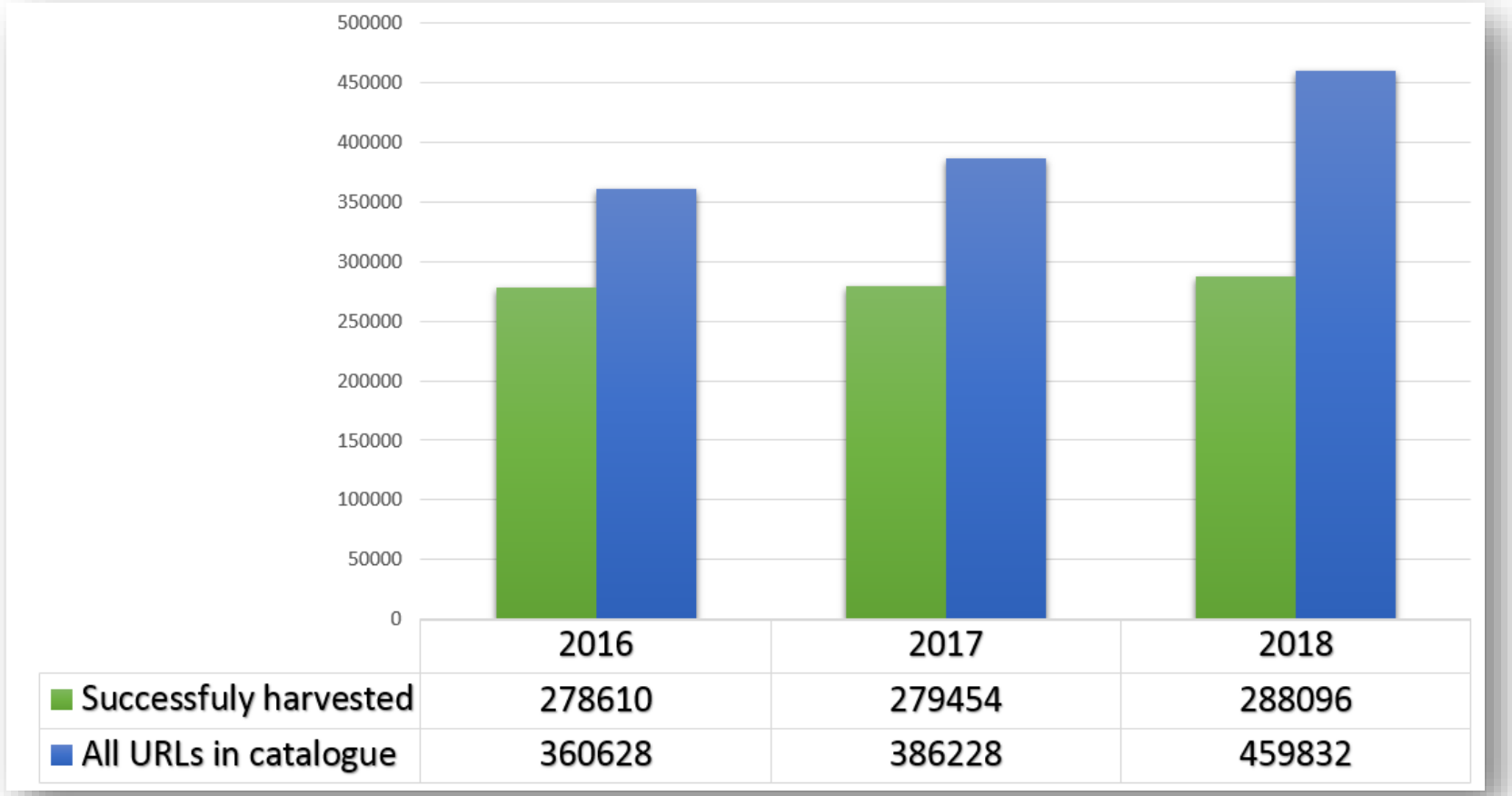
HARVEST TYPES - 2: THEMATIC HARVEST

- Thematic (Event) harvest
- 1 GB, 2 days/ domain
- Olympic Games, elections, Tour de France, ...
- respecting robots.txt

Good morning Cannondale Club [photograph]. In: Pinterest [online]. [cit. 2018-10-10] Available on: https://sk.pinterest.com/CdaleClub/peter-sagan

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.



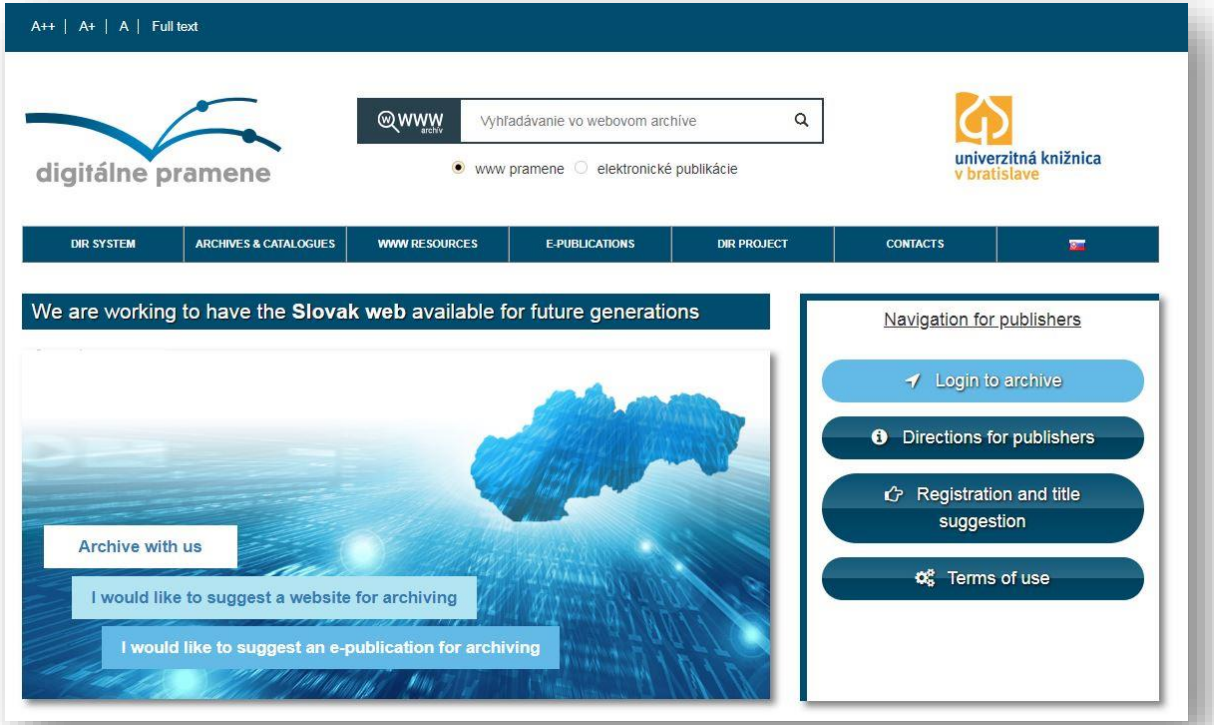
	2016	2017	2018
Successfuly harvested	278610	279454	288096
All URLs in catalogue	360628	386228	459832

HARVEST TYPES – 3: FULL-DOMAIN HARVEST

Information about full-domain harvests per year	2016	2017	2018
Number of nominations (proposed websites)	360 628	386 228	459 832
Number od nominations successfully collected	278 610	279 454	288 096
Compressed size	6.85 TB	6.35 TB	6.18 TB
Uncompressed size	10.5 TB	17.2 TB	9.33 TB
Number of WARC files (max. 2 GB/WARC)	278 663	232 747	288 266

PUBLIC PORTAL

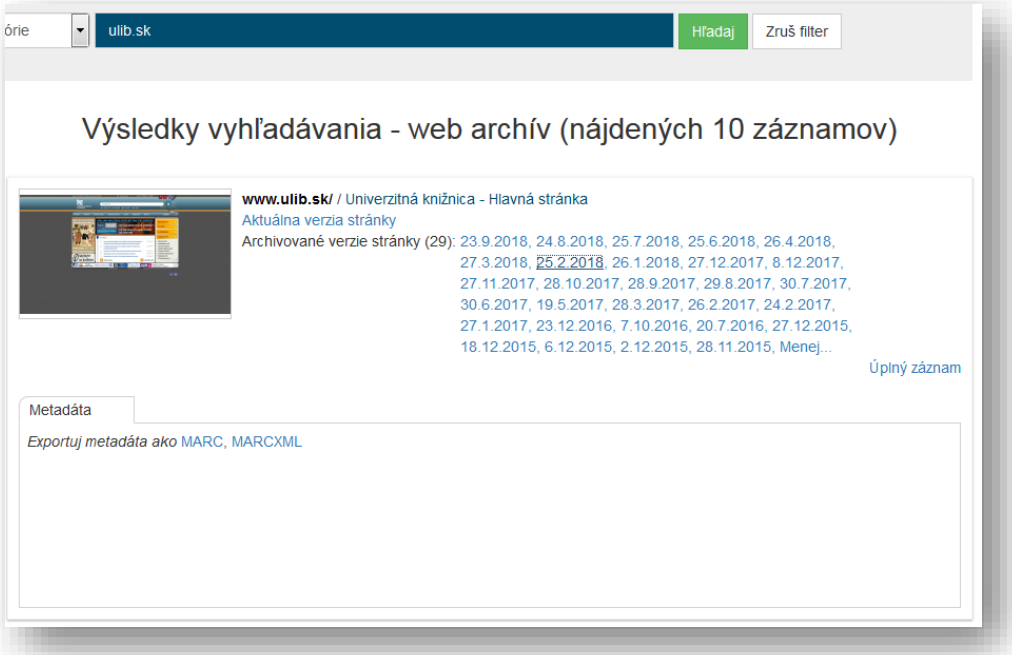
- a lot of information for publishers including forms for publication submission/ suggestion
- frontend for web archive and e-publications archive
- plenty of information about our activities
- made in Wordpress
- simple and advanced search



https://webdepozit.sk

SIMPLE SEARCH

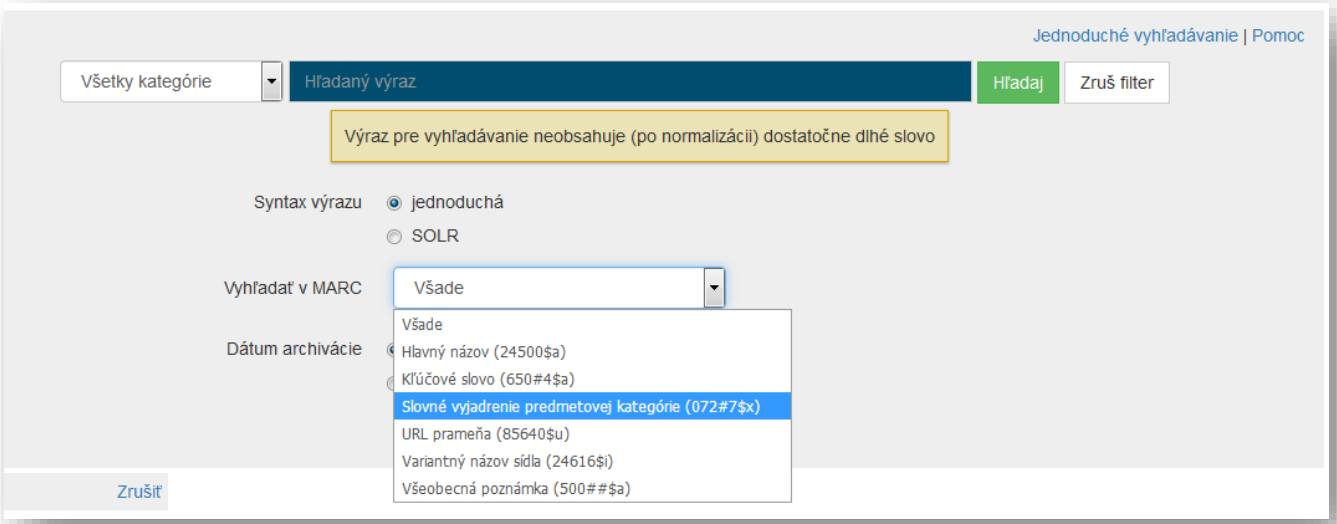
- search in all metadata fields
- filtered by Conspectus categories
- links to live websites and archived versions

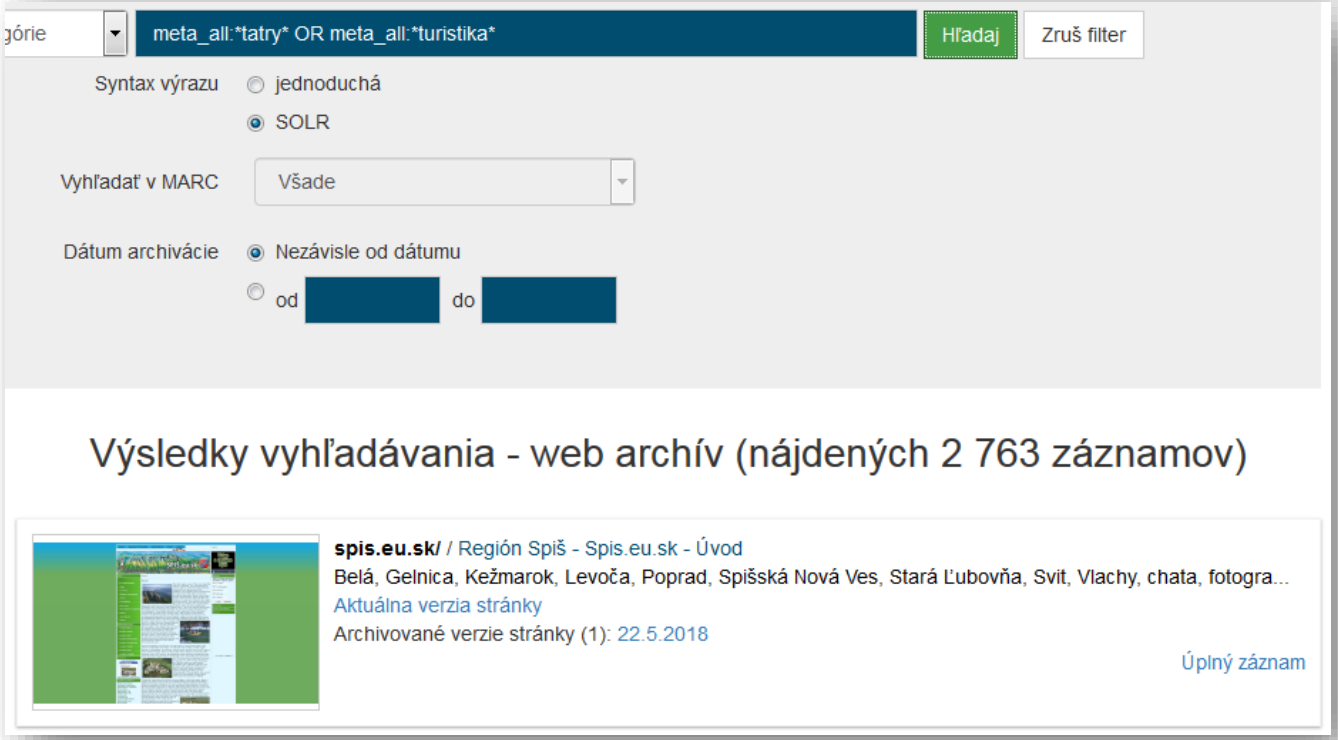


Search of the keyword „ulib.sk“ – links to live site, OW versions, possible export of harvest metadata

ADVANCED SEARCH

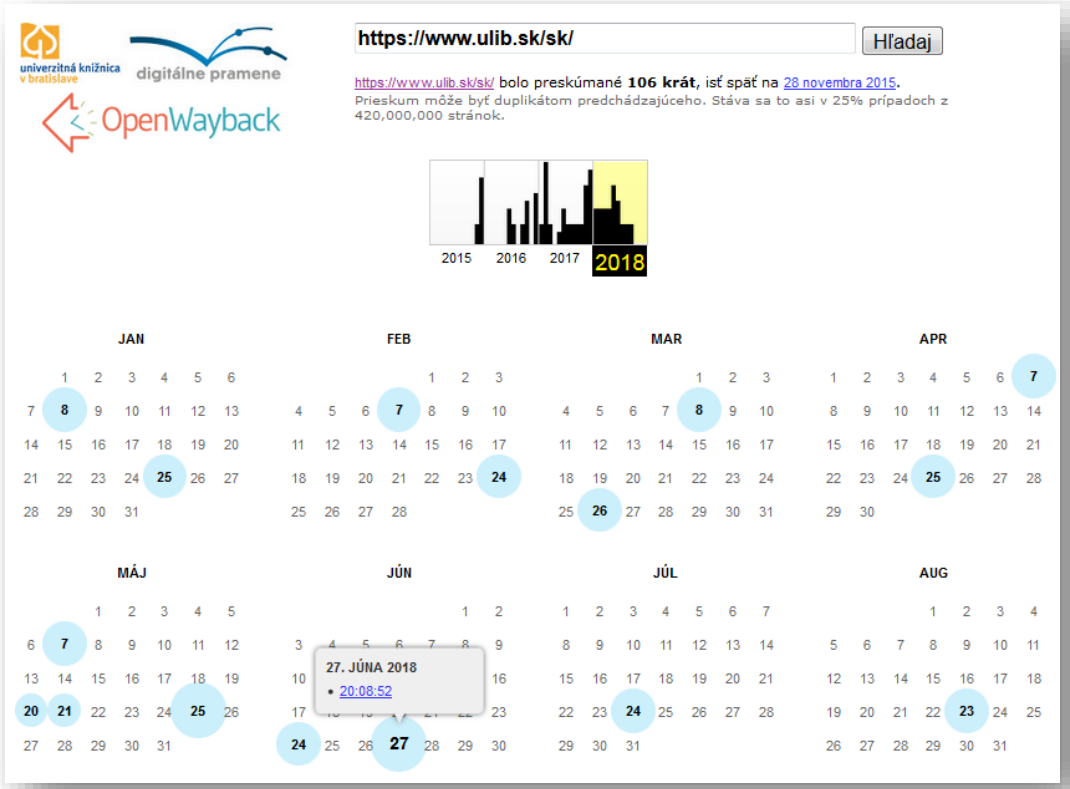
- Basic metadata search by following MARC fields:
  - All metadata
  - Title (24500$a)
  - Keyword (650#4$a)
  - Conspectus category (072#7$x)
  - URL of resource (85640$u)
  - Subtitle (24616$i)
  - Description (500##$a)
- SOLR complex expression search using operators
  - Logical operators:
    - and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





OPENWAYBACK

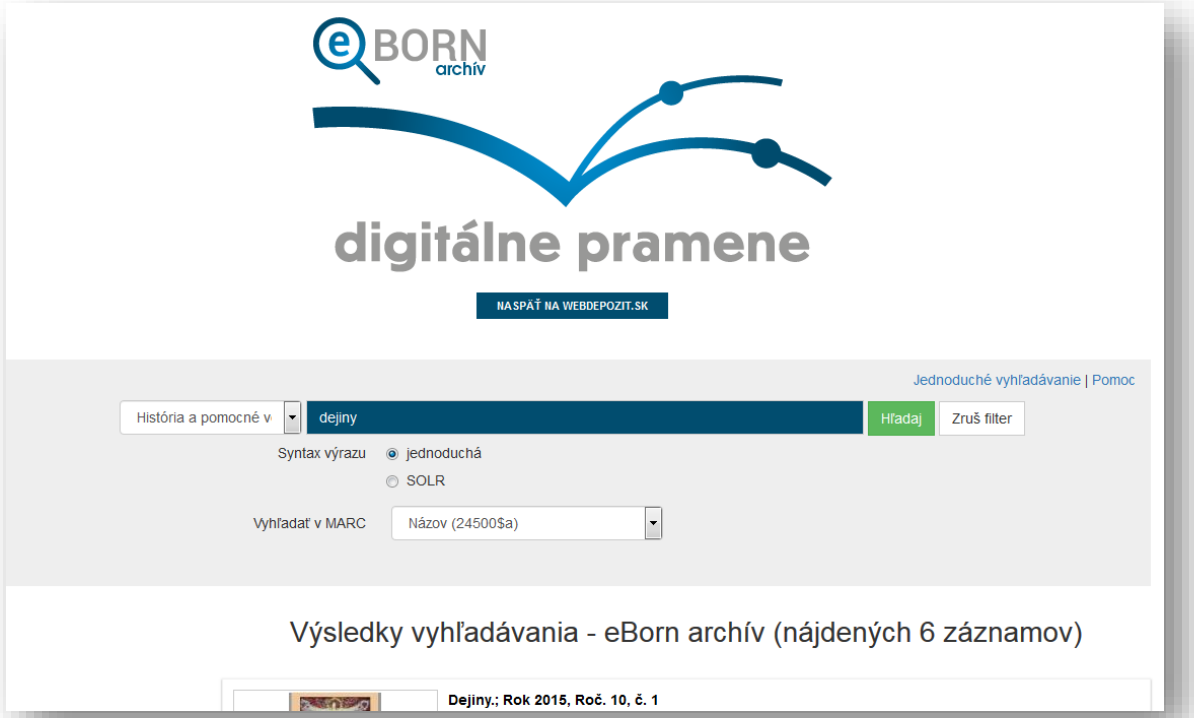
- Results in OpenWayback
- Displaying of the archived website

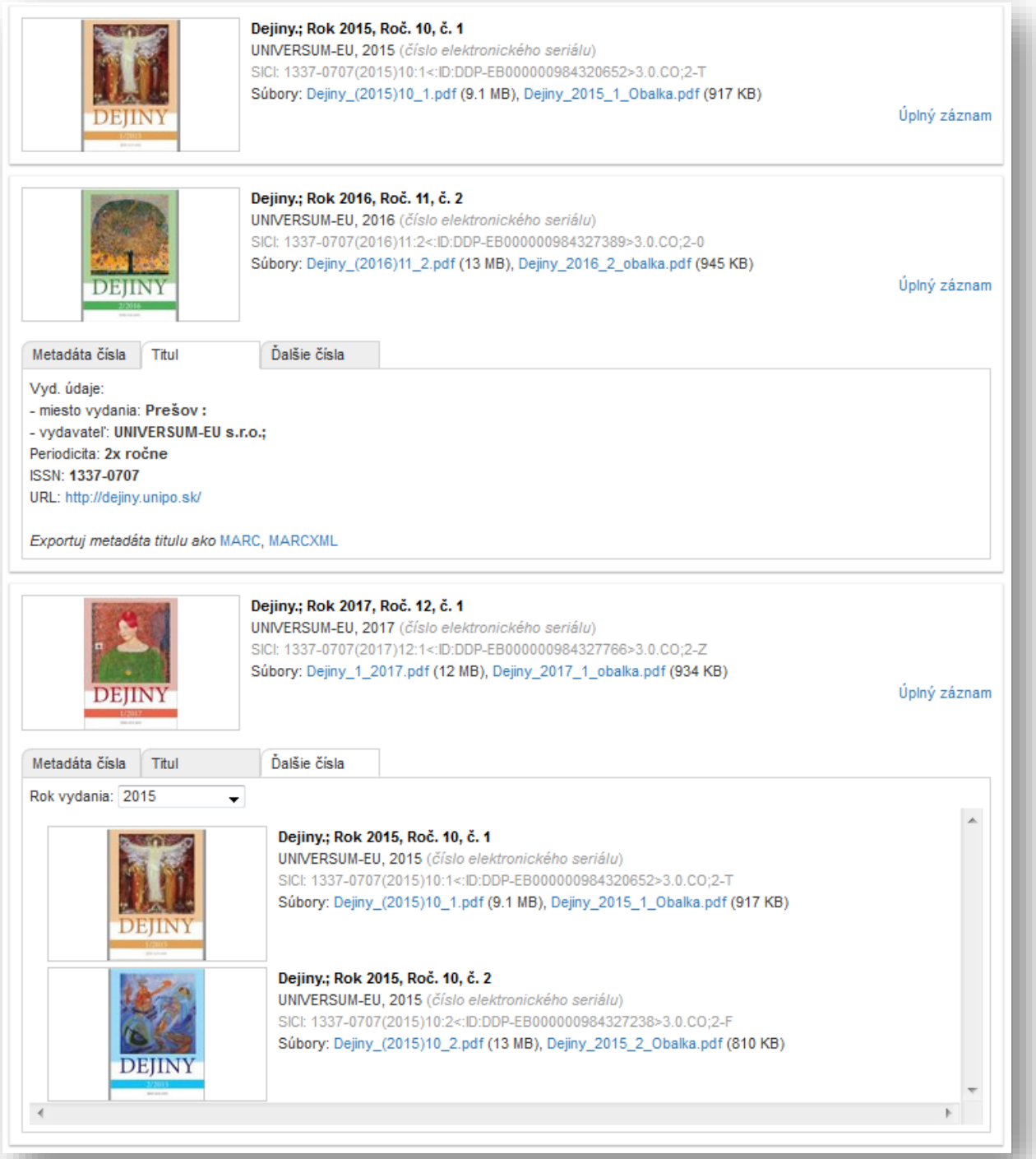


https://search.webdepozit.sk

E-BORN SEARCH

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





ACCEPTED FORMATS

FORMAT IDENTIFIERS:

- DROID

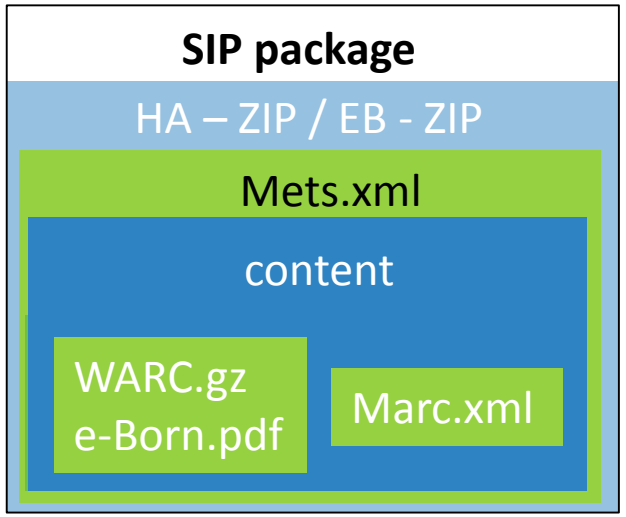
VALIDATORS:

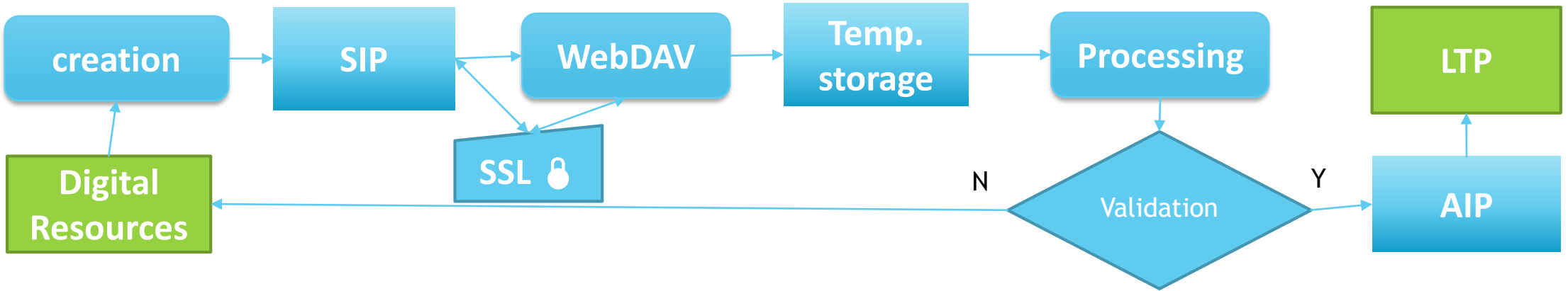
- PDF - JHOVE
- PDF/A - veraPDF
- WARC - JHOVE
- ePub – epubcheck

NAME;	MIME TYPE	PUID/ID	MIME TYPE
PDF (1.0)	application	fmt/14	application/pdf
PDF (1.1)	application	fmt/15	application/pdf
PDF (1.2)	application	fmt/16	application/pdf
PDF (1.3)	application	fmt/17	application/pdf
PDF (1.4)	application	fmt/18	application/pdf
PDF	application	fmt/19	application/pdf/1.5, application/pdf/undef
PDF (1.6)	application	fmt/20	application/pdf
PDF (1.7)	application	fmt/276	application/pdf
PDF/A (1a)	application	fmt/95	application/pdf
PDF/A (1b)	application	fmt/354	application/pdf
PDF/A (2a)	application	fmt/476	application/pdf
PDF/A (2b)	application	fmt/477	application/pdf
WARC	application	fmt/289	application/warc
ePub	application	fmt/483	application/epub+zip

LONG-TERM PRESERVATION OF THE SLOVAK ARCHIVE OF DIGITAL RESOURCES

- Digital Resources - content of the archive is sent into the LTP repository (Central Data Archive)
- SIPs are compressed
- SIP types: HA – harvest & EB – e-Born
  - HA – mets.xml, WARC – marc.xml + warc.gz
  - EB – mets.xml, PDF – marc.xml + name.pdf





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

OPIS OPERAČNÝ PROGRAM INFORMATIZÁCIA SPOLOČNOSTI

TVORÍME VEDOMOSTNÚ SPOLOČNOSŤ

univerzitná knižnica v bratislave