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
UNIVERSITY LIBRARY IN BRATISLAVA

Michalská street
Franz Liszt’s Pavilion and Garden
Ventúrska st. – courtyard with glass ceiling
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

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

- Web sites on the internet
  - SK domains
  - Other domains
- Digital publications
  - Source of DP
  - Automated jobs DP
  - Heritrix web harvester
  - Antivirus check
  - Deduplicator, Web Content Analyzer
- LTP archive (CDA)
  - CDA Location A
  - CDA Location B
- Public portal
  - Open Wayback
  - DP upload
  - DP view
  - Search
  - OA-PAH provider
- Deposit of digital publications (DDP)
  - Archive of digital publications (files)
  - Archive of web sites (WANG)
  - Database DP
- Internal web app
  - Web app based on WebCurator tool functionality
  - Invenio
- SIP upload
  - DIP upload
  - Catalogue DP
  - Index builder
  - Metadata DP
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 overview
- 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)

Domain status in the catalogue
- available - 323 331 (67%)
- unavailable - 151 123 (31%)
- unknown - 8 108 (2%)
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

<table>
<thead>
<tr>
<th></th>
<th>2015</th>
<th>2016</th>
<th>2017</th>
<th>2018</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selective</td>
<td>1</td>
<td>7</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>Thematic</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Full-domain</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SUM</td>
<td>3</td>
<td>12</td>
<td>14</td>
<td>18</td>
</tr>
</tbody>
</table>

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
HARVEST TYPES - 2: THEMATIC HARVEST

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

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.

<table>
<thead>
<tr>
<th>Year</th>
<th>Successfully harvested</th>
<th>All URLs in catalogue</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016</td>
<td>278610</td>
<td>360628</td>
</tr>
<tr>
<td>2017</td>
<td>279454</td>
<td>386228</td>
</tr>
<tr>
<td>2018</td>
<td>288096</td>
<td>459832</td>
</tr>
</tbody>
</table>
## HARVEST TYPES - 3: FULL-DOMAIN HARVEST

<table>
<thead>
<tr>
<th>Information about full-domain harvests per year</th>
<th>2016</th>
<th>2017</th>
<th>2018</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of nominations (proposed websites)</td>
<td>360 628</td>
<td>386 228</td>
<td>459 832</td>
</tr>
<tr>
<td>Number of nominations successfully collected</td>
<td>278 610</td>
<td>279 454</td>
<td>288 096</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Compressed size</th>
<th>6.85 TB</th>
<th>6.35 TB</th>
<th>6.18 TB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncompressed size</td>
<td>10.5 TB</td>
<td>17.2 TB</td>
<td>9.33 TB</td>
</tr>
</tbody>
</table>

| 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 Wordpres
- 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
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

![Diagram of the process]

**SIP package**
- HA – ZIP / EB - ZIP
- Mets.xml
- content
- WARC.gz
- e-Born.pdf
- Marc.xml

**Digital Resources**

**SSL**

**WebDAV**

**Temp. storage**

**Processing**

**Validation**

**N**

**Y**

**LTP**

**AIP**
THANK YOU FOR YOUR ATTENTION

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